Media Sentiments for Enhanced Credit Risk Assessment

Jin-Chuan Duan* and Xuan Yao†

(First Draft: April 13, 2022; This Version: August 22, 2022)

Abstract

We deploy a combination of natural language processing (NLP) techniques (Source-LDA, NER and TABSA-BERT) to extract credit-focused, entity-specific media sentiments on all North American public firms (US and Canadian) over recent 20 years. Around 1.7 million English news articles collected from three reputable business presses form the corpus. Our objective is to treat media sentiments as alternative data and study their potential in complementing the already rich structured financial variables (common risk factors and individual attributes). Our findings based on a three-category (default, other exit and survival) classification reveal that media sentiments add significant explanatory power to default prediction on those firms with credit-relevant media coverage. In addition, the mere fact of getting media coverage, regardless of being positive or negative, can help predict corporate exits due to reasons other than default (mainly mergers and acquisitions).

Keywords: Default, NLP, LDA, BERT, Sentiment Analysis, Logistic Regression

*Duan is with National University of Singapore (Business School and Asian Institute of Digital Finance). E-mail address: bizdjc@nus.edu.sg.

†Yao is with National University of Singapore (Asian Institute of Digital Finance). E-mail address: yaoxuan@nus.edu.sg

1 Introduction

Text in business press have long been playing an important role in business decision making. Official filings motivated by disclosure of vital business activities and analyst reports constitute other sources of corporate information in the text form. These traditionally text-based information has largely been digitized and ready for machine-based analyses. These text streams in principle represent collective and systematic expressions that are likely of value to decision making. In contrast to numerical data, text streams are unstructured and more challenging to handle (see Chan and Chong; 2017). The impacts of media sentiment on financial analysis for such as stock return, volatility and credit risk have been studied but are still in the early phase of a growing literature; for example, Groß-Klußmann et al. (2019), Sun et al. (2020), Xing et al. (2019), Roeder et al. (2020), Tsai et al. (2016) and Dunham and Garcia (2021).

This paper distinguishes in the way of extracting media sentiments that are topic-focused, entity-specific and at the article level as opposed to generic sentiments expressed on firms. To reflect the credit focus of this study, we devise a way to determine an article's extent of relevance to credit risk and then use it to weight the sentiments expressed on all firms mentioned in the article. The sentiments expressed on firms are naturally sentence-based, but we will aggregate them to the article level. Sentiments expressed on different firms appearing in the same article can also vary so that our derived sentiment scores are entity-specific. We favor business press over social media because the opinions expressed by professional journalists are arguably more insightful and objective, and the editorial control helps ensure quality and consistency (Yu et al.; 2013, Li et al.; 2020). Our study utilizes daily business press instead of other types of text because they form systematic and timely information sources (Tsai et al.; 2016).

To guide our natural language processing (NLP) work specifically for enhancing credit risk analysis, we must first understand the nature of corporate default prediction. In a nutshell, it is a supervised learning that links firm-by-firm categorical outcome, say, Y not yet realized to some predictive attributes $\boldsymbol{X}$ available at the time of prediction. The Altman (1968) Z score is an early and probably most cited academic work on the topic of credit risk analysis. Advancements in the area have naturally brought to the literature many sophisticated modeling approaches (econometric/statistical and machine learning) and helped identify technical issues of real relevance. On the use of logistic regressions, Shumway (2001) has for example pointed out an inherent dynamic dependency in the time series of corporate data that must be respected for estimation and statistical inference. It is relevant to our study because we will deploy logistic regression, a classical classification tool, to demonstrate the use of media sentiments to enhance credit risk analysis. In addition, we shall factor in an often overlooked aspect of corporate credit analysis, and that is, corporate exits may be due to reasons other than default, for example, merger and acquisition.

Other forms of corporate exit along with default determine a firm's survival and its importance cannot be understated. In the literature, factoring in other exits in default prediction has already

been implemented, for example, in the models of Duffie et al. (2007) and Duan et al. (2012), among others. Worth noting is the fact that the other-exit occurrence rate is typically an order of magnitude larger than the default rate. Duan et al. (2012) has, for example, reported such evidence on US public firms. The North American sample used in our study also exhibits the same property. Beyond affecting survival probability, predicting other exits can have its own implications and financial significance. In short, the categorical outcome variable for default analysis should at least be of three classes so that Y can take on the value of, say, 0, 1 or 2 to respectively denote survival, default or other exit.

To determine the extent of an article's relevance to credit risk, we must recognize that credit risk is not a narrowly focused and clearly defined matters such as camera, surf board, etc. which can be captured by a simple word or phrase. To understand the complexity of capturing an article's degree of relevance to a topic, we first need to appreciate the power and limitations of a modern sentenced-based alternative, i.e., the targeted aspect-based sentiment analysis (TABSA-BERT) of Saeidi et al. (2016). TABSA-BERT is a fine-tuning approach relying on BERT of Devlin et al. (2018), which requires of lumping together target (corporate name) and aspect (topic) as a composite entity so that it can used to create an auxiliary sentence to complement the original sentence and through which to link to the expressed sentence-based sentiment. Since our topic of interest, credit risk, cannot be easily captured by a simple word, phrase or sentence, it should not treated as a sentence-based matter and best left to an article-level holistic assessment like what humans would do. We deploy Source-LDA of Wood (2016) for an article-level topic extraction on a corpus of around 1.7 million English articles gathered from three reputable business presses (Financial Times, Thomson Reuters, and Wall Street Journal). We opt for Source-LDA instead of LDA of Blei et al. (2003) because our need to guide the extraction of the credit risk topic's word distribution and to determine an article's degree of relevance to the defined topic.

However, we still deploy TABSA-BERT for sentence-based sentiment analysis where the aspect is treated as absent and a target (corporate name) is replaced with a generic location identifier depending on its sequence of appearance in the sentence when multiple corporate names appear in a sentence. This ensures that the supervised-learning model for sentiment assignments can be generically applicable to all corporate entities. The TABSA-BERT method is a sentence-pair approach with an auxiliary sentence specifically created for each entity in a sentence to accommodate potentially different sentiments expressed on more than one entity. The method essentially couples BERT with another simple feed-forward network that links the BERT output corresponding to a sentence pair to an annotated sentiment. We apply the TABSA-BERT method to train the sentiment assignment model with a five-category ordinal scale on a sample of three thousand or so sentences in 600 randomly selected articles with each meeting the 30% credit risk relevance.

Being able to identify corporate and other names is essential to our topic-focused, entity-specific sentiment analysis. We deploy NLP tools including Named-Entity Recognition (NER) of spaCy (Honnibal and Montani; 2017) and Coreference Resolution of Stanza (Qi et al.; 2020) to identify

corporate names and their demonstrative pronouns in articles. With them, we can implement a remove-insert approach for topic extraction. Removal of corporate and other names is performed before conducting topic extraction because we want the credit risk topic to be as generic as possible and is not tied to any corporate and other types of names. Corporate names are also needed to complete the task of substituting them with location identifiers for the TABSA-BERT supervised-learning task. Although our interest in this study is limited to a set of North American corporate names that we already have, we still need to identify other corporate names that may appear in articles for the removal and/or substitution operation.

We construct a sentiment assignment model by applying a five-category ordinal scale on the entire corpus and use it to obtain entity-specific, sentence-based sentiments. These sentiments will be aggregated up to entity-specific at the article-level sentiment scores, and each of which is further weighted by an article's relevance to credit risk captured by the Source-LDA analysis. On a daily basis, we average sentiment scores across multiple articles from the same media, and then further average them across the three business presses. Furthermore, we run seven-day moving averages to obtain final daily time series of credit-focused, entity-specific at the article-level media sentiments.

We add these time series of media sentiments to the structured financial variables on the 17,296 North American public firms (US and Canadian) available in the NUS-CRI database.[1] We take over 20-year monthly time series of this North American component of the NUS-CRI dataset which contains corporate events (default and other exit) along with a set of predictive variables (17 macro-financial factors and individual firm attributes) that have already been deployed in the NUS-CRI model. The NUS-CRI model is based on the forward-intensity approach of Duan et al. (2012) and has achieved the accuracy ratios of 94.1% and 85.7% for one-month and one-year predictions, among other horizons up to five years.[2] For our purpose and to avoid the complex issues associated with modeling term structure of default, we use a simple three-class logistic regression and focus on commonly considered one-year prediction. This allows us to answer the central question of whether media sentiments can enhance prediction performance.

Two media sentiment variables are created with one being our credit-focused, entity-specific at the article-level media sentiment score on a daily basis and the other being a dummy variable indicating whether a sentiment score is available for a particular date. Then, we take the month-end data sentiment score and dummy to match the final firm-month observations of the structured financial variables. Our empirical findings reveal that media sentiments add statistically significant explanatory power to default prediction on those firms with credit-relevant media coverage. Interestingly, the sentiment dummy is highly significant in predicting other exit. Thus, the mere fact of getting media coverage, regardless of being positive or negative, can help predict the other-exit events which are mainly mergers and acquisitions.

[1] Please refer to NUS-CRI (2021a) for general information and the data description. The web link is `https://nuscri.org/en/`.

[2] Please refer to Table B.1 on page 119 of NUS-CRI (2021a).

The remainder of this paper is organized as follows. Section 2 discusses various methods critical to this study. Section 3 describes the data source and sample construction. Section 4 presents the empirical findings and finally Section 5 concludes the paper with a discussion on implications, limitations and future research.

2 Methodology

For corporate default prediction, we deploy a three-class logistic regression which combines the structural financial variables and media sentiments (score and dummy). A proper use of logistic regression for corporate default analysis is not trivial and requires some elaborations. We also rely on a combination of NLP techniques (Source-LDA, NER and TABSA-BERT) to extract the credit-focused, entity-specific media sentiments for which we will provide discussions in different subsections.

2.1 Default prediction with a three-class logistic regression

Our default prediction uses a simple three-class logistic regression on the the North American (US and Canadian) firms in the NUS-CRI database (NUS-CRI; 2021a). The dependent variable is categorical taking on the value of 0, 1 and 2 (representing survival, default and other corporate exit in the subsequent year). The predictive feature variables are 17 structured financial variables (common risk factors and individual attributes) plus the two alternative data fields – sentiment score and sentiment dummy. In total, we face 19 feature variables giving rise to 40 parameters with 20 each default and other exit (adding an intercept).

Because only a relatively small subset (64,670) out of a large number of total firm-month observations (over 1.5 million) has media sentiment scores, we face a critical task of parameter estimation with a very large number of missing values. Originally, our media sentence-based sentiment scores are recorded as -2, -1, 0, 1 and 2 with 0 standing for a neutral opinion. After converting to an article-level score, multiplying by an article's relevance to the credit risk topic, and further aggregations (over articles of the same media source, across three business presses and over a seven-day moving window), our credit-focused, entity-specific median sentiment scores are values bounded between -2 and 2 but are not necessarily integers.

For cases of missing media sentiment score, we assign zero. But we also recognize that a zero sentiment score generated by the language model is conceptually not equivalent to the zero due to the missing value assignment. Intuition suggests that our sentiment score may yield a bias reflective of the typical media coverage towards news-worthy stories. Indeed, our data confirms this intuition with an average sentence-level sentiment score at -0.028. From a technical angle, the sentiment dummy should be viewed as an offset to the potential distortion arising from the missing data treatment. Looking at this from an economic analysis, the sentiment dummy can reflect whether

the mere fact of getting media coverage can have an impact on the chance of occurrence of default and/or other exit.

Conducting a logistic regression for corporate default prediction requires some careful thought as pointed out by Shumway (2001). Some notation are helpful in making the issue clear. Let $1_{\{y_{it}(\tau)=j\}}$ be an indicator function taking on the value of 1 if $y_{it}(\tau) = j$ ($j = 0, 1$ or 2) and 0 otherwise. $y_{it}(\tau)$'s value is determined by the corporate event of the i^{th} firm (out of n firms) taking place in the period between t and $t+\tau$, for example, viewing τ as one year. $p_j(\boldsymbol{x}_{it}; \boldsymbol{\theta}, \tau)$ is a three-class logistic function determining the probability of class-j's occurrence in $(t, t+\tau]$. Each firm-i has a time index set, $\mathcal{T}_i$, containing all available time points that the τ-period default prediction can be made. The log-likelihood function of the sample (or the negative cross-entropy loss function using the machine-learning language) can be expressed as

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Y}_{1:n}, \boldsymbol{X}_{1:n}, \tau) = \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \tau) = \sum_{i=1}^{n} \sum_{t \in \mathcal{T}_i} 1_{\{y_{it}(\tau)=j\}} \log p_j(\boldsymbol{x}_{it}; \boldsymbol{\theta}, \tau). \tag{1}$$

Shumway (2001) made two important observations. First, $\mathcal{L}_i(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \tau)$ should be viewed as a single quantity because it duly reflects the dynamic structure of survival up to a time point in conjunction with the subsequent corporate event. A static-model treatment that misses a firm's prior survival up the prediction time point would produce inconsistent parameter estimates. Therefore, a logistic regression should include all previous time points to duly reflect their survival probabilities such as $\mathcal{L}_i(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \tau)$ and it is then equivalent to a survival analysis. His second point is concerned with the sample size in computing statistics. In short, the sample should be viewed as n observations with $\mathcal{L}_i(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \tau)$ as one composite quantity for each firm.

In the original formulation of Shumway (2001), however, a two-class logistic regression was used and his treatment thus overlooked corporate exits due to reasons other than default, which distorts survival probabilities. Both Table B.1 of Duan et al. (2012) and the results reported later in this paper indicate that the chance for other exits to occur is an order of magnitude larger than default probability. This omission is likely to have serious consequences. In short, one should at the minimum use a three-class instead of two-class logistic regression in modeling corporate defaults.

In addition to the above discussions, we face an added issue arising from overlapping structure of default predictions. Our data runs on a monthly frequency but the prediction is for the year that immediately follows. Although this overlapping structure is rooted in the natural real-world setting, it requires an appropriate treatment. In a nutshell, $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Y}_{1:n}, \boldsymbol{X}_{1:n}, \tau)$ is only a pseudo log-likelihood function due to the overlapping structure; that is, two consecutive one-year default predictions overlap by eleven months. The form of the objective function stated above with a single firm relies on conditional independence. It is obviously untrue and thus it is not a true log-likelihood function. We can obviously remove the overlapping structure by dropping the monthly time series to opt for annual one-year default predictions. But that would result in the loss of a huge quantity

of information conveyed in media sentiments.

Our solution resorts to Duan et al. (2012) which shows that using the pseudo log-likelihood due to overlapping periods continues to preserve consistency but the standard errors need robustification. In short, we deploy $\mathcal{L}_i(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \tau)$ as a composite quantity and use the typical sandwich estimator of the covariance matrix for the parameter estimates; that is, combining the Hessian matrix of the whole sample with the outer product of individual gradient vectors. Specifically, the pseudo maximum likelihood estimator, $\hat{\boldsymbol{\theta}}$, has the following robustified asymptotic variance:

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\theta}}) &= \left(\frac{\partial^2 \mathcal{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{Y}_{1:n}, \boldsymbol{X}_{1:n}, \tau)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1} \sum_{i=1}^{n} \left(\frac{\partial \mathcal{L}_i(\hat{\boldsymbol{\theta}}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \tau)}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \mathcal{L}_i(\hat{\boldsymbol{\theta}}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \tau}{\partial \boldsymbol{\theta}} \right)' \\ &\quad \times \left(\frac{\partial^2 \mathcal{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{Y}_{1:n}, \boldsymbol{X}_{1:n}, \tau)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1}. \end{aligned} \tag{2}$$

The variance suggested by Shumway (2001) can be understood as equal to $\left(\frac{\partial^2 \mathcal{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{Y}_{1:n}, \boldsymbol{X}_{1:n}, \tau)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1}$. It is of course the outcome of the Fisher information theory when the prediction periods are not overlapped.

2.2 Identify credit-focused, entity-specific at the article-level sentiments

Three NLP components are central to our extraction of sentiments from the business media corpus. After extraction, we need to aggregate sentiments across articles, different business presses, and over a moving time window. Our aim is to produce time series of credit-focused, entity-specific at the article-level media sentiments to serve as alternative data. They will then be used to complement those structured financial variables in the three-class logistic regression model to examine whether there is incremental explanatory power.

2.2.1 Named Entity Recognition (NER)

We need to identify corporate and other names in the business media corpus for three reasons. First, removal of corporate and other names allows us to form generic credit risk topic distribution that is free of those entity names. Second, we also need to remove corporate names from sentences so that the sentiment assignment model can be trained in a way that is applicable to all firms. Finally, we need to match corporate names after insertion back into articles with those corporate names in the NUS-CRI database on North American public firms so that media sentiments can be pooled together with the structured financial variables.

NER is an NLP technique to detect and classify entities (e.g., companies, persons, locations, etc) in a text document. We use Python-based spaCy (v3), an NLP package providing various practical tools for text processing, including NER to identify corporate and other names. We further fine-tune with 10,000 sentences the standard NER model that is based on the spaCy's

English pipeline. These 10,000 sentences are randomly taken from our corpus with each sentence containing at least a corporate name on the benchmark list, which is the NUS-CRI database of North American corporate names. The choice of this benchmark list reflects the fact that these firms are the ultimate target of interest for this study.

In addition to recognizing company names, it is important to consolidate their different forms for the same company that often appear in media; for example, International Business Machines and IBM. The trailing sentence after a sentence with a corporate name may use a pronounce to describe the same firm. It is therefore critical that we are able to associate it with the right company. To handle both situations, we resort to Coreference Resolution of Stanza (Qi et al.; 2020) to find reference to the same corporate entity in a pair of adjacent sentences in the same article where the first sentence contains a corporate name.

We calculate the similarity scores between an entity in articles and the company names on the benchmark list. Once the similarity score exceeds the threshold value of 0.9, we view them as the same company. The practical mapping details are more complicated and go through several rule-based pre-processing steps to improve accuracy and efficiency; for example, we drop Ltd, Corp, etc from a company name before computing the similarity score.

2.2.2 Extracting credit risk topic with Source-LDA

This paper utilises Source-LDA (Wood; 2016) to extract the credit risk topic's word distribution from a business media corpus. Source-LDA is a variant of LDA (Latent Dirichlet allocation) of Blei et al. (2003). LDA is a bag-of-words approach by viewing each article as a combination of a fixed set of topics weighted by the article-specific probabilities. Each topic is in turn defined by a word distribution which is common to all articles. LDA finds a best representation of articles in a corpus by MCMC utilizing a collapsed Gibbs sampler.

Deploying Source-LDA reflects our specific need to generate the credit risk topic. Guiding the topic with a prior distribution holds the key to ensuring that a reasonable credit risk topic distribution can emerge. For our purpose, we only set two topics and the credit risk topic is the one that is given a strong prior word distribution. The second topic has no specific prior distribution to guide its formation and is meant to be a generic catch-all device. An article's probability on the credit risk topic attributed by Source-LDA is the weight that we use to determine its degree of relevance to credit risk. Different from Wood (2016)'s approach of relying on Wikipedia for the prior knowledge in Source-LDA, we select 20 articles directly from our media sources that are deemed highly relevant to credit risk.

It is worth noting the standard Bayesian convergence result and its implication here. The Walker (1969) Theorem implies that the prior belief's impact on the credit risk topic distribution

will dissipate with a growing corpus. Hence, we cap the corpus size at 50,000 articles for topic-extraction and tunes those hyper parameters to achieve our objective. According to Wood (2016), the LDA model shows convergence after 1,000 iterations.

To implement Source-LDA properly, we first remove company and other types of names (by NER), punctuation, and only keep the nouns and verbs by Part-of-Speech (POS) tagging, and then convert to lowercase and lemmatize. Moreover, we identify those two adjacent words (bigrams) forming a specific meaning and commonly seen in discussions of credit risk. We connect them with a underline to make it a new word, for example, interest_rate, credit_risk. In addition, stopwords and sentiment words are removed (using the sentiment words dictionary of Loughran and McDonald (2015)).

2.2.3 Sentiment analysis by TABSA-BERT

BERT (Bidirectional Encoder Representations from Transformers) of Devlin et al. (2018), a pre-trained language model, has in recent years gained popularity and delivered promising results for sentiment analysis. Utilizing a transformer architecture of Vaswani et al. (2017), BERT has simplified though a pre-trained language model the text-based supervised-learning tasks. Its design facilitates an easier implementation of downstream NLP tasks such as classification through fine-tuning the pre-trained model's parameters along with the training of an add-on layer of supervised-learning neural network. This flexibility can more effectively cater to unique text aspects inherently important to different supervised learning tasks.

A sentence is a sequence of words: $\{w_1, \ldots, w_m\}$. The input text to BERT can be either a single sentence and a sentence pair. Figure 1 provides a diagram for the classification task with a sentence-pair input. The input to BERT (the row below the box) is a sequence of representations of (sub)word tokens forming the sentence pair and the BERT's output (the top row in the box) is a set of encoding corresponding to words in the input sentence pair. BERT has two versions – BERT-base and BERT-large, encoding the tokens with a vector length of 768 and 1,024, respectively. We deploy in this paper RoBERTa-large developed by (Liu et al.; 2019), an improved version of the BERT model. Encoding reflects word, segment and position embedding. In short, the same word can have different encoding to reflect its specific meaning in a sentence; for example, the word 'bank' in commercial bank vs river bank. Two unique tags, [CLS] and [SEP] as shown in Figure 1, are the sentence-level embedding and a sentence separator, respectively. [CLS] stands for classification whose encoding at the output layer is meant to summarize the sentence pair and serve as an encoded input to the subsequent classification task.

Since different sentiments may be directed at different targets, the sentence-pair approach basically creates an auxiliary pseudo sentence for a target to couple with the original sentence. For example, two targets in a sentence will result in two sentence-pairs with each being tagged by a target-specific sentiment. In Sun et al. (2019)'s TABSA-BERT which is a targeted aspect-based

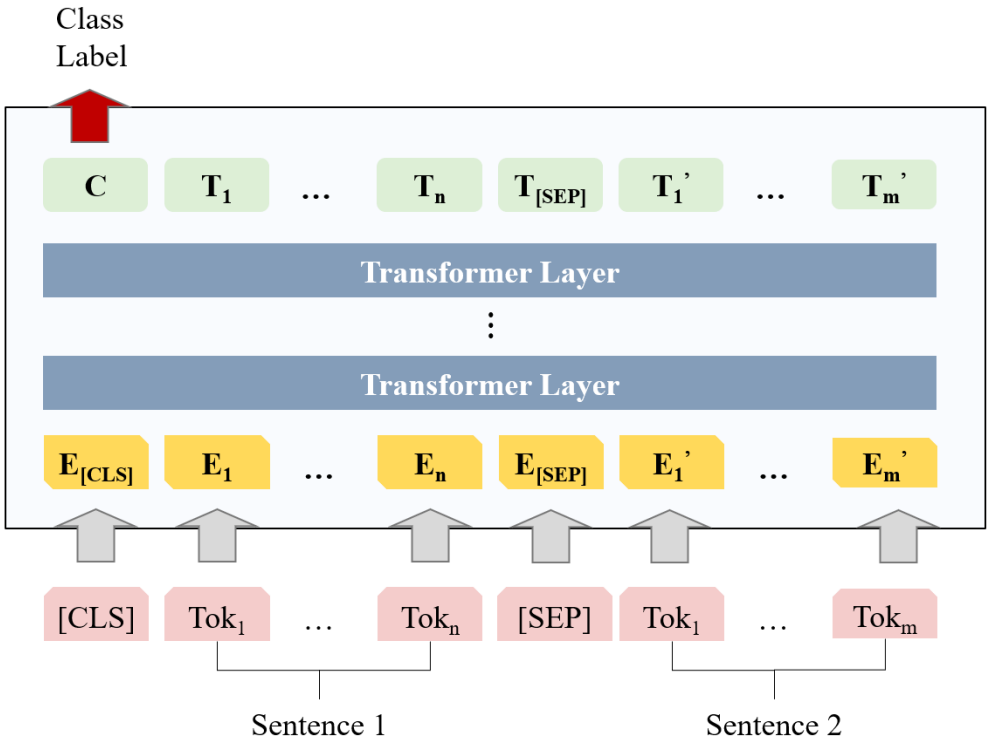


Figure 1: Sentence-pair fine-tuning with BERT

sentiment analysis utilizing BERT, one needs to identify target(s) and the aspect of interest and associated them with sentiment(s). Targets are corporate names in our case, but the task has no specific aspect to deal with because our earlier discussion already concludes that relevance to credit risk should be holistically assessed at the article level and left for Source-LDA.

The sentence-pair approach for our 5-class ordinal classification task is best illustrated with the example in Table 1. 'Target1' and 'Target2' are generic words introduced to recognize the companies and their positions in a sentence. Since our design does not rely on specifying an aspect, the second sentence in the sentence-pair only specifies a target. Were the word 'credit risk' the aspect, we would simply create four sentence-pairs to accommodate four combinations. In addition, we would need to revise the four second sentences to 'Target1-credit risk', 'Target2-credit risk', 'Target1-other' and 'Target2-other' and tag the last two auxiliary pseudo sentences with a sentiment of 'none' to reflect the fact that their sentiments are irrelevant to the aspect.

Because our sentiment labels on a five point scale from -2 to 2 are ordinal, it makes sense to perform ordinal classification instead of categorical classification to be more line with the linguistic understanding. When conducting BERT fine-tuning classification, 'C' corresponding to [CLS] is the input vector for the add-on layer of classification neural network. Since we use RoBERTa-large, 'C' is a 1024-dimensional vector of real numbers between 0 and 1. In short, one can view RoBERTa-large as converting a sentence or in our case a sentence pair into 1024 features. Our ordinal classification network has one hidden layer with four nodes and takes 'C' as the input. After

Table 1: An example of TABSA-BERT task

The original sentence:
Compared to Boeing's mountain of debt, analysts say Airbus appears much more comfortable, sitting on $7 billion in net cash and a second-mover's advantage.
The modified sentence:
Compared to Target1's mountain of debt, analysts say Target2 appears much more comfortable, sitting on $7 billion in net cash and a second-mover's advantage.
Sentence-pair #1 tagged by a label of -2 (strongly negative)
1.1: The modified sentence
1.2: Target1
Sentence-pair #2 tagged by a label of +2 (strongly positive)
2.1: The modified sentence
2.2: Target2

averaging these four node values, we invert it with the logistic function and apply four cutoff values to obtain the five probabilities by the logistic function to correspond to the five ordinal classes. This classification neural network has 4×1025 biases and synaptic weights linking 'C' to the four hidden nodes and another 4 cutoff values to define the five ordinal classes.

2.2.4 Sentiment aggregation

Entity-specific sentiments are at the sentence level and they need to be converted to article-level sentiments. For each article-level sentiment, individual article's credit risk topic weight can be applied to obtain credit-focused, entity-specific at the article-level sentiment. Further aggregations to the single-media level, across multiple media sources and over time are still needed to become a default prediction feature.

Specifically, we go through the following aggregation steps:

1. Identify a relevant sentence block for a company in an article by starting with the first sentence with that company's name. Add to the sentence block with the trailing sentence if it contains the same company name. The trailing sentence is still added to the block if Coreference Resolution of Qi et al. (2020) identifies the pronoun or abbreviation to represent the same firm. Continues to add to the block until exhausting all sentences in an article.

2. Each sentence in the sentence block has its own sentence-level sentiment score allocated by the sentiment assignment model. We take the sentence-level sentiment score with the maximum magnitude as our entity-specific, article-level sentiment score. Using a maximum-magnitude sentence-level sentiment instead of a block-average reflects a belief that the punchline sentence is more reflective of the sentiment expressed on a company in an article.

3. The credit-focused, entity-specific at the article-level sentiment is generated by multiplying the article-level, entity-specific sentiment with the article's credit risk topic weight.

4. On a given day, one business press may have multiple articles covering the same company. We thus average the credit-focused, entity-specific at the article-level sentiment scores by the same media source. A further equally-weighted average is applied over the business presses. Equal weights are deemed appropriate in our case because the three media sources are all highly respectable.

5. News coverage may not be available on consecutive days, but news may still have lingering impacts. We thus set a moving window of seven calendar days and compute the seven-day moving average for credit-focused, entity-specific sentiments to produce the daily time series for each media-covered company. Missing sentiment scores will not be substituted. The moving average will reflect the number of non-missing sentiment scores in the seven-day window. Since our default prediction study runs on a monthly frequency, we in essence uses monthly snapshots of the daily time series.

3 Data and variables

3.1 Corporate events

The corporate events of default and other exits are obtained from the Credit Research Initiative (CRI) database at the National University of Singapore (NUS) (NUS-CRI; 2021a). The NUS-CRI team collects credit events from numerous sources, including Bloomberg, Compustat, Moody's and other credit-rating-agency reports, exchange websites, and news sources for over eighty thousand exchange-listed firms around the world.

Our sample has 17,296 US and Canadian public firms (both financial and non-financial) over the period from June 1998 to June 2021 on a monthly frequency. There are altogether 1,580,842 firm-month observations in this sample. Among them, there are 14,796 firm-month default observations (less than 1%) and 121,394 firm-month other-exit observations (7.68%). Note that one default, for example, may generate up to 12 firm-month default observations due to conducting one-year default predictions on a monthly frequency.

3.2 Business media corpus

We construct the sentiment database by merging three business presses: Financial Times, Thomson Reuters, and Wall Street Journal. Our media corpus contains 1,726,376 articles between May 1998 and June 2021. The ending month reflects the fact that it is the final month at which time the one-year default prediction can be checked against the corporate actions realized in the following year (running up to June 2022). Table 2 shows the downloadable article counts and the time span for each media source. Figure 2 reports the count distribution over time.

After mapping the company names with those having structured financial variables in our sample, there are 5,904 firms with media coverage out of the total 17,296 North American public

Table 2: Media sources (articles in "Sports" or "Life & Arts" excluded)

Media Source	Counts	Start Date	End Date
Financial Times	142,475	2004-08-02	2021-06-30
Thomson Reuters	545,989	2006-10-24	2021-06-30
Wall Street Journal	1,037,912	1998-05-31	2021-06-30

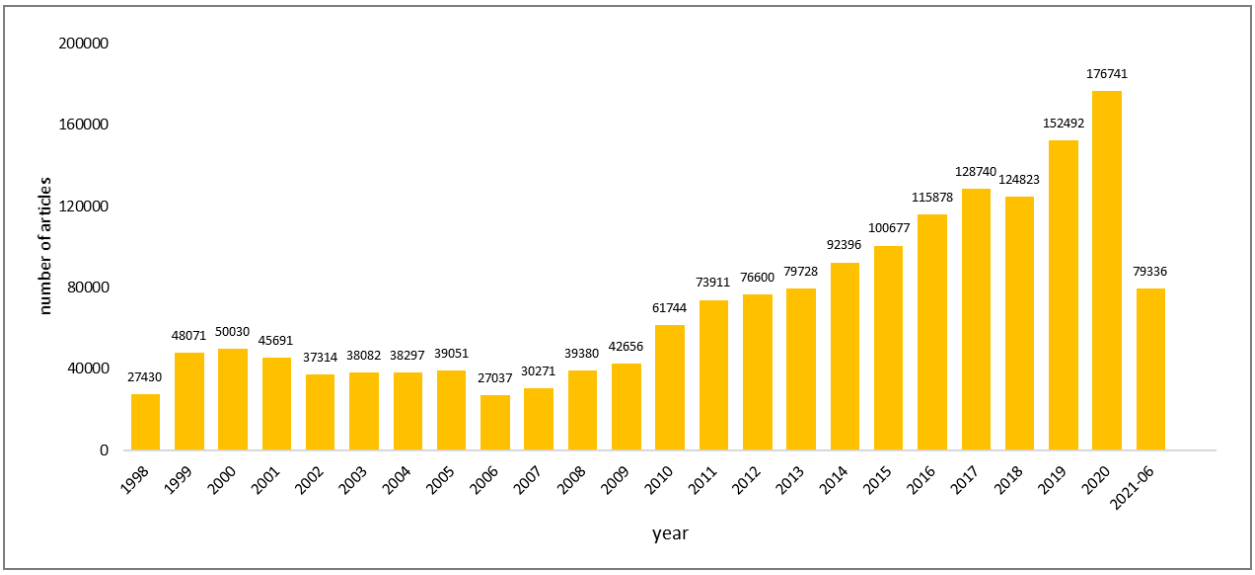


Figure 2: Article numbers over the sample years

firms at month ends over the sample period. The total news article count is 710,405 containing 3,080,589 sentences. Although about one-third of the companies have been covered by the three business presses, but many have low appearance frequency, resulting in a significant number of missing data cases for which we have previously discussed in Section 2.1 and have devised a way to handle them statistically.

3.3 Credit-focused entity-specific media sentiments

3.3.1 Topic's word distribution and article's topic weight distribution

We randomly select 50,000 articles as the training dataset and apply the hyper parameters governing the Dirichlet prior distribution on the credit risk topic (see Appendix A). Source-LDA produces the word distributions on the credit risk and the other catch-all topics. Figure 3 is a visual presentation of word clouds for the two topics. The word size in the cloud represents the order of importance for a topic-word distribution.

Figure 4 provides credit risk topic's weight distribution for all news from the three media sources and those news covering our sample of public companies in the US and Canada. It is evident that only a small set of articles are highly related to credit risk, which is consistent with one's intuition.

(a) Credit risk topic

(b) Catch-all topic

Figure 3: Word clouds for the credit risk and catch-all topics

When it is narrowed down to the news linked to North American public companies, a relatively larger proportion of news are highly credit risk related.

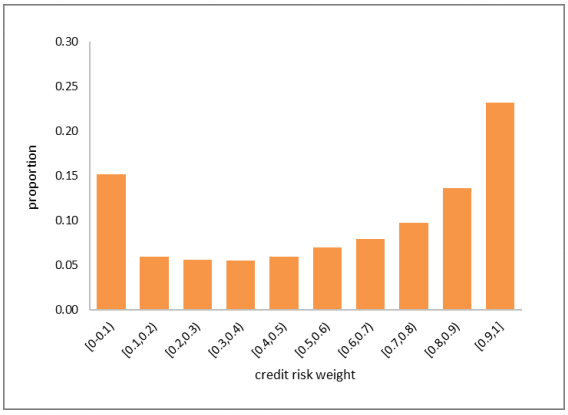
3.3.2 Entity-specific sentiments

Table 3 provides summary statistics on the sentiment distribution at the sentence level for the 14,488 of 17,296 US and Canadian public firms that had coverage of at least one article over the sample period by any one of the three business presses. We produce these results with the implementation details given in Appendix B. Clearly indicated is the fact that most sentence-level sentiments are neutral. The median is 0 but the mean equal to -0.028 is slightly negative, suggesting an overall negative opinion expressed by these three business presses. The number of strongly negative sentiments is fifty percent more than that of strongly positive opinions. This is hardly surprising because strongly negative news are news worthy.

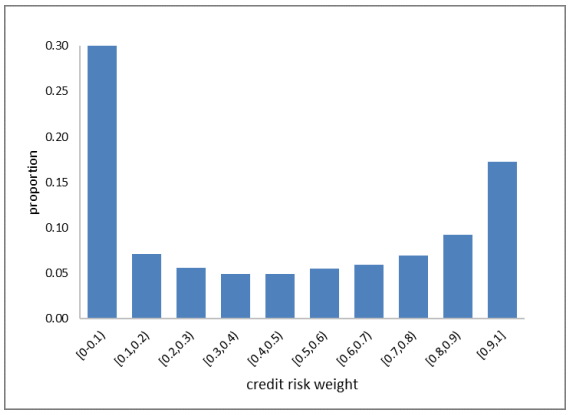
Table 3: Sentence-level sentiment distribution

Total Count	Count of 0	Count of -2	Count of 2	Mean	Median	Std
3,080,589	2,020,779	129,147	85,994	-0.028	0.000	0.743

Figure 5 shows company's sentiment counts at the sentence level. There are 10,092 out of 14,488 companies registering equal or less than 25 sentence-sentiments. These sentence-level sentiment counts imply that news tends to focus on a small group of companies. This narrow focus reflects an expected selection bias towards news worthiness, and the pattern can again be seen at the article-level sentiment counts. A similar big tail occurs at the article level (see Figure 6).



(a) Articles covering the public companies in the US and Canada



(b) All articles

Figure 4: Credit risk weight distributions

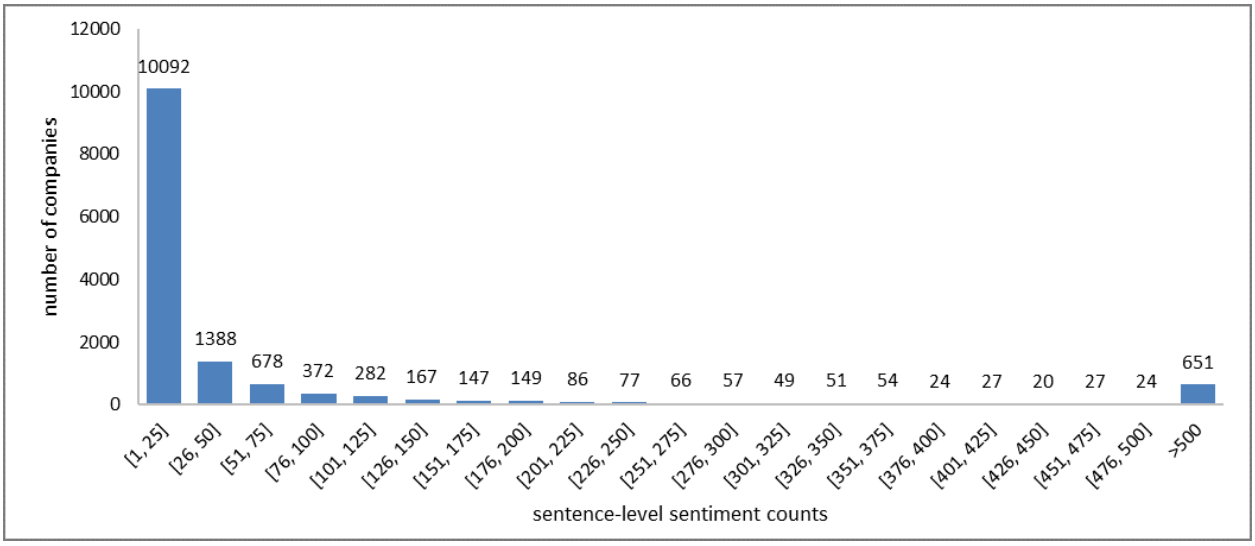


Figure 5: Company's sentence-level sentiment counts distribution

3.4 Structured financial variables (control variables)

Structured financial variables used in default prediction studies are many. In this study, we adopt the set of default predictors currently deployed in the live NUS-CRI corporate default prediction system on its North American sample (NUS-CRI; 2021a). The choice of these structured financial variables was largely motivated by Duan et al. (2012). These variables fall in two large categories: (1) macro-financial risk factors and (2) firm-specific attributes. Some are constructed from market prices/rates whereas others are based on quarterly financial statements. A dummy variable is also included to distinguish financial from non-financial firms. There are in total 17 structured financial variables to serve as the control variables in our examination of the enhancement role of media sentiment for default prediction.

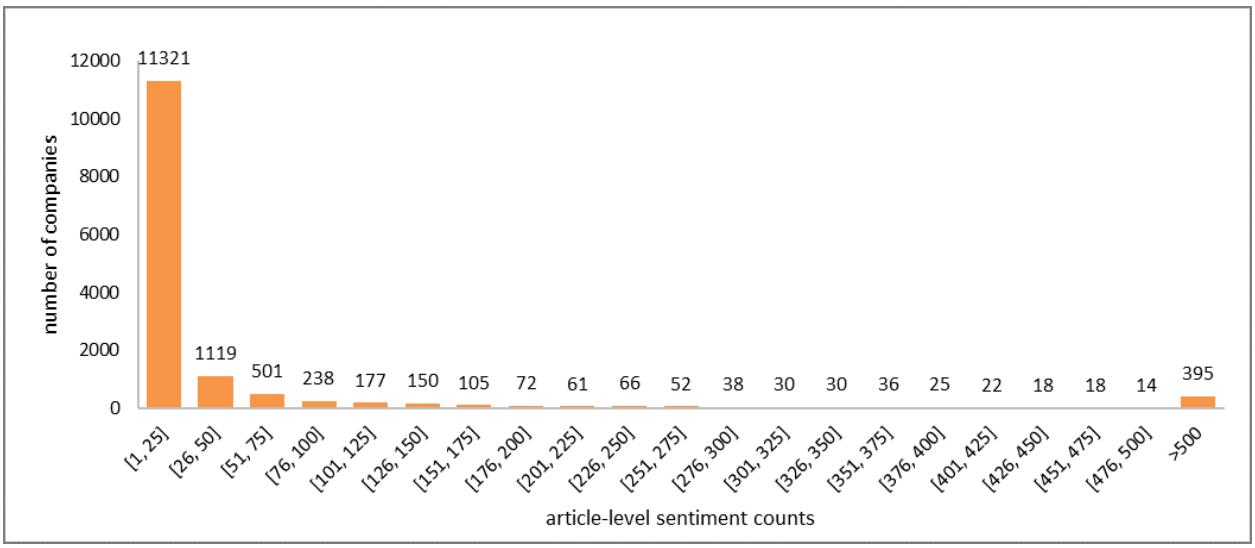


Figure 6: Company's article-level sentiment counts distribution

Macro-financial risk factors

1. Stock Index Return: the trailing one-year simple return on a major stock index of the economy;
2. Interest Rate: a representative 3-month short-term interest rate standardized from the data availability point;
3. Financial Aggregate DTD: median distance-to-default of financial firms in each economy inclusive of those foreign firms whose primary stock exchange is in this economy;
4. Non-financial Aggregate DTD: median distance-to-default of non-financial firms in each economy inclusive of those foreign firms whose primary stock exchange is in this economy.

Firm-specific attributes (13 firm-specific attributes including five transformed measures on some of the following firm characteristics)

1. DTD: the estimated firm-level distance-to-default as a measure of volatility-adjusted leverage;
2. CASH/TA: a log ratio of cash and short-term investments to total assets as a measure of liquidity for financial firms;
3. CA/CL: a log ratio of current assets to current liabilities as a measure of liquidity for non-financial firms;
4. NI/TA: a ratio of net income to total assets as a measure of profitability;
5. SIZE: a log ratio of market capitalization to the economy's median market capitalization as a measure of relative size;
6. M/B: a ratio of the sum of market capitalization and total liabilities to total assets as a measure of market mis-valuation/future growth opportunities;
7. SIGMA: the standard deviation of the residuals of a regression of the daily returns of the firm's market capitalization on the daily returns of the economy's stock index as a measure of idiosyncratic volatility;

8. FIN Dummy: a dummy variable to indicate a financial firm.

The first five characteristics are also transformed into level and trend measures where the level measure is computed as the one-year moving average whereas the trend is calculated as its current value minus the one-year moving average. Duan et al. (2012) and the subsequent NUS-CRI implementation (NUS-CRI; 2021a) have found the trend measure to add significant predictive power for short-term prediction horizons.

4 Empirical findings

4.1 Statistical findings

Table 4 reports the summary statistics of all the predictive features used in this study (sentiment-related and structured financial variables). One may suspect that the two newly introduced sentiment variables are correlated with some of the structured financial variables. The correlations involving the sentiment variables are provided in Table 5, which reports the correlation coefficients between them and the firm-specific control attributes. In general, sentiment score and sentiment dummy have very weak correlations in magnitude with the structured financial variables. The only exception is the correlation of 0.27 between sentiment dummy and size level, suggesting that bigger firms get more media attention but the expressed sentiment may be positive or negative.

Table 4: Descriptive statistics for the feature variables

Variables	Mean	STD	Min	25%	Median	75%	Max
Sentiment Variables							
Sentiment Score	0.000	0.1315	-2	0	0	0	2
Sentiment Dummy	0.041	0.198	0	0	0	0	1
Structured Financial Variables							
Stock Index	0.070	0.175	-0.546	-0.015	0.097	0.173	0.713
Interest Rate	-0.35	0.760	-1.182	-1.068	-0.566	0.336	1.393
DTD Level	4.041	3.489	-1.121	2.048	3.451	5.301	78.694
DTD Trend	-0.087	1.258	-15.758	-0.620	-0.044	0.508	7.598
CA/CL Level	0.629	0.867	-4.012	0	0.484	1.1	4.709
CA/CL Trend	-0.017	0.313	-2.483	-0.083	0	0.056	2.619
NI/TA Level	-0.007	0.046	-1.085	-0.004	0.001	0.005	0.207
NI/TA Trend	0	0.034	-0.534	-0.002	0	0.002	0.537
Size Level	0.221	2.063	-6.139	-1.247	0.136	1.586	6.697
Size Trend	-0.03	0.353	-1.906	-0.182	-0.02	0.135	1.997
M/B	1.679	3.425	0.156	0.769	1.002	1.595	83.951
Sigma	0.171	0.124	0.017	0.086	0.135	0.216	1.092
CASH/TA Level	-0.672	1.446	-9.653	0	0	0	0
CASH/TA Trend	-0.001	0.236	-3.36	0	0	0	3.314
AggDTD for Fin	0.65	1.377	0	0	0	0	5.794
AggDTD for nonFin	2.81	1.701	0	2.035	3.001	4.254	5.458
FIN Dummy	0.208	0.406	0	0	0	0	1

Table 5: Correlations of sentiment variables with others

Firm-specific Variables	Sentiment Score	Sentiment Dummy
Sentiment Variables		
Sentiment Score	1.0***	0.008***
Sentiment Dummy	0.008***	1.0***
Structured Financial Variables		
DTD Level	0.014***	0.069***
DTD Trend	0.017***	-0.001*
CA/CL Level	0.002***	-0.051***
CA/CL Trend	0.002***	0.007***
NI/TA Level	0.002***	0.041***
NI/TA Trend	0.003***	-0.001
Size Level	0.003***	0.271***
Size Trend	0.039***	0.011***
M/B	0.009***	0.008***
Sigma	-0.005***	-0.092***
CASH/TA Level	0.000	0.027***
CASH/TA Trend	-0.001	0.003**
FIN Dummy	0.001	-0.012***

Note: Statistical significance levels: *p <0.10, **p <0.05, ***p <0.01.

Negligible correlations suggest that information embedded in the sentiment variables is unlikely carried by the structured financial variables. But this does not automatically imply that these sentiment variables convey useful predictive information. The three-class logistic regression for predicting one-year default and other corporate exit can help shed light on this central issue. For default prediction, Table 6 shows that the credit-focused, entity-specific media sentiment is statistically significant at the 1% level, but the sentiment dummy is not based on any typical significance level. The negative coefficient turns out to be as expected, implying that a negative sentiment score increases a company's one-year default probability. The statistical significance used here is computed with the robust standard error that we have already discussed in Section 2.1.

When it comes to the other exit event, both sentiment variables are highly significant, again using the robust standard errors. Their coefficients are positive, and the sentiment dummy seems to play an even stronger role. It suggests that being reported, regardless of being a positive or negative sentiment, can increase a company's one-year probability of other exit. The sentiment score acts to complement the impact further if the media coverage expresses a positive sentiment. Note that the majority of other corporate exits are due to mergers and acquisitions. Thus, our empirical findings on the probability of other exit appear to be intuitively plausible because a financially-distressed firm often becomes a cheap acquisition target but a well-performing company may be even more attractive to potential suitors.

Table 6: The 3-class logistic regression for default and other exit

	Default		Other Exit	
	Coefficient	Std Err (robust)	Coefficient	Std Err (robust)
Sentiment Variables				
Sentiment Score	-0.182***	0.062 (0.069)	0.213***	0.023 (0.025)
Sentiment Dummy	0.083	0.055 (0.107)	0.431***	0.017 (0.038)
Structured Financial Variables				
Intercept	-3.447***	0.054 (0.150)	-2.471***	0.018 (0.058)
Stock Index	0.031	0.054 (0.159)	0.181***	0.020 (0.054)
Interest Rate	0.233***	0.013 (0.038)	0.285***	0.004 (0.013)
DTD Level	-1.337***	0.011 (0.042)	0.019***	0.001 (0.003)
DTD Trend	-0.940***	0.018 (0.047)	0.073***	0.003 (0.008)
CA/CL Level	-0.045	0.012 (0.036)	-0.173***	0.004 (0.013)
CA/CL Trend	-0.354***	0.021 (0.046)	-0.239***	0.009 (0.019)
NI/TA Level	-0.658*	0.162 (0.381)	-0.553**	0.067 (0.206)
NI/TA Trend	-1.199***	0.142 (1.95)	-1.152***	0.073 (0.129)
Size Level	0.153***	0.006 (0.018)	-0.168***	0.002 (0.007)
Size Trend	-1.227***	0.021 (0.049)	-0.460***	0.008 (0.022)
M/B	-0.032***	0.005 (0.008)	-0.031***	0.001 (0.005)
Sigma	-0.101	0.081 (0.223)	1.879***	0.029 (0.103)
CASH/TA Level	0.031	0.024 (0.076)	-0.096***	0.005 (0.019)
CASH/TA Trend	0.105	0.051 (0.129)	0.035	0.013 (0.027)
AggDTD for Fin	0.146	0.029 (0.090)	-0.022	0.006 (0.020)
AggDTD for nonFin	0.402***	0.013 (0.037)	-0.04***	0.004 (0.014)
FIN Dummy	-0.58	0.116 (0.359)	-0.459***	0.029 (0.097)
# of Observations	1,580,842			

Note: Statistical significance uses the robust standard error: *p <0.10, **p <0.05, ***p <0.01.

4.2 Economic impacts on PD and POE

Our sample has 64,670 observations with a sentiment score out of the total 1,580,842 firm-month data instances. The sentiment variables are therefore unlikely to materially alter the coefficients on the structural financial variables and their impacts on PDs and POEs are expected to surface only on the sub-sample. For the examination of whether inclusion of the sentiment variables exerts impact of economic significance, we thus re-estimate the model with only structured financial variables and calculate another set of PDs and POEs.

Figure 7 exhibits the distribution of PD changes for this sub-sample of 64,670 data instances due to the introduction of two media sentiment variables. First, more than a half of the data points have a higher PD. Second, most data points have rather small PD changes whereas about one tenth of the sample see changes in excess of 10 basis points (bps). To gain a better practical perspective on 10 bps, we note that agency-reported, one-year historically realized default rates for BBB-rated firms typically hover around 15 bps.

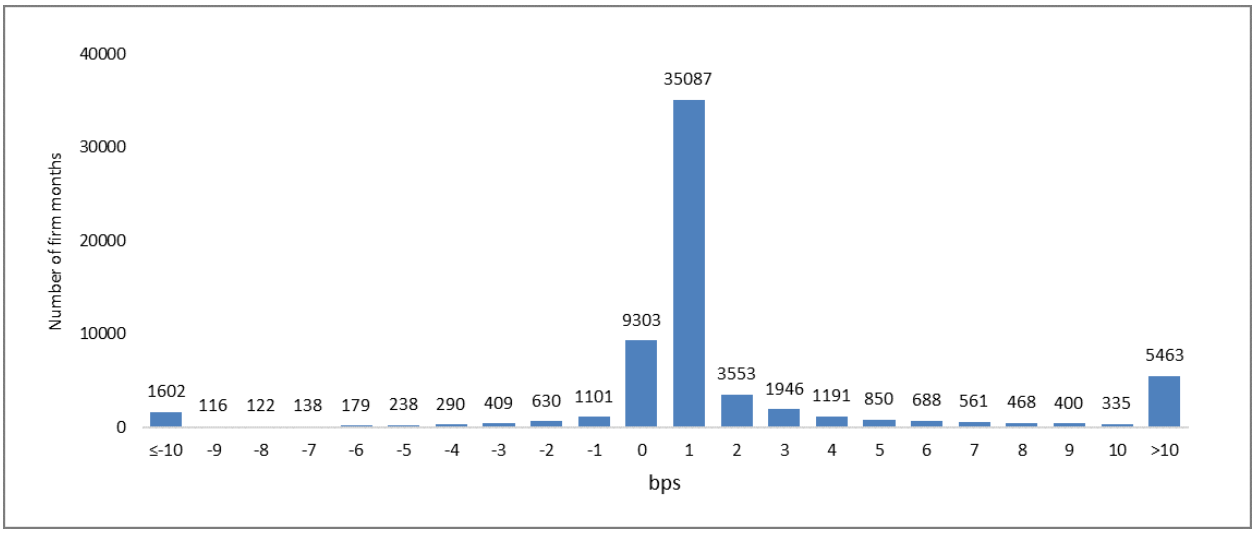


Figure 7: The distribution of PD changes due to the inclusion of media sentiments

Relying on agency credit ratings is a commonly accepted business practice. The device of PD-implied rating, which converts a PD to an implied credit rating, can help readers appreciate the practical impact of a PD change resulting from the inclusion of media sentiments in default prediction. We use the PDiR2.0 method of Duan and Li (2021), which is a mapping design to translate the one-year PD into an equivalent credit rating by referencing the S&P or Moody's historical credit migration experience. For our purpose, we use the PDiR2.0 that references the S&P credit migration, which has already been implemented in the NUS-CRI operation (NUS-CRI; 2021b). The impact on the PD-implied ratings are displayed in Figure 8, omitting the cases of no change. Evident from the 8,346 implied rating changes (at least one notch) out of 64,670 firm-month data instances with sentiment scores, downgrading by one notch far exceeds an upgrade of one notch with 5,799 vs 2,278. From a business practice perspective, this suggests that incorporating media

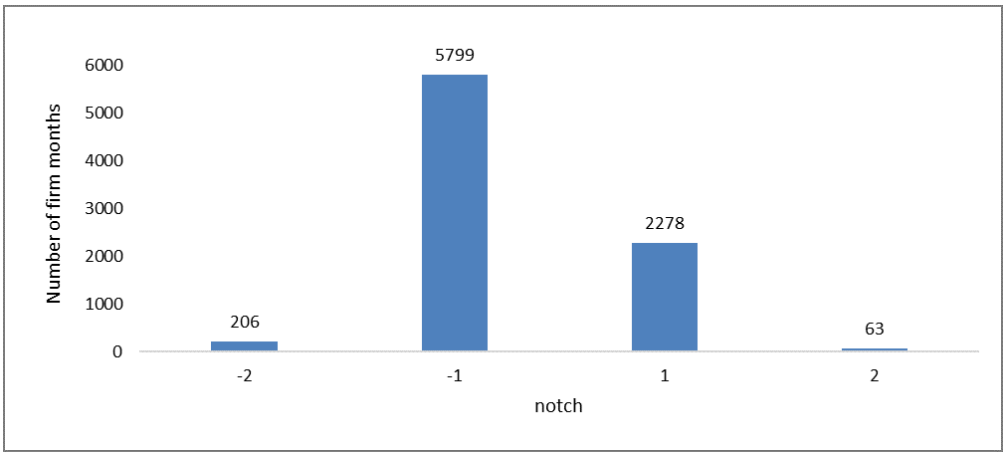


Figure 8: The PD-implied credit rating changes (in notches) due to the inclusion of media sentiments (cases of no change have been omitted)

sentiments has materially pushed up the overall assessment of credit risk for these North American public firms.

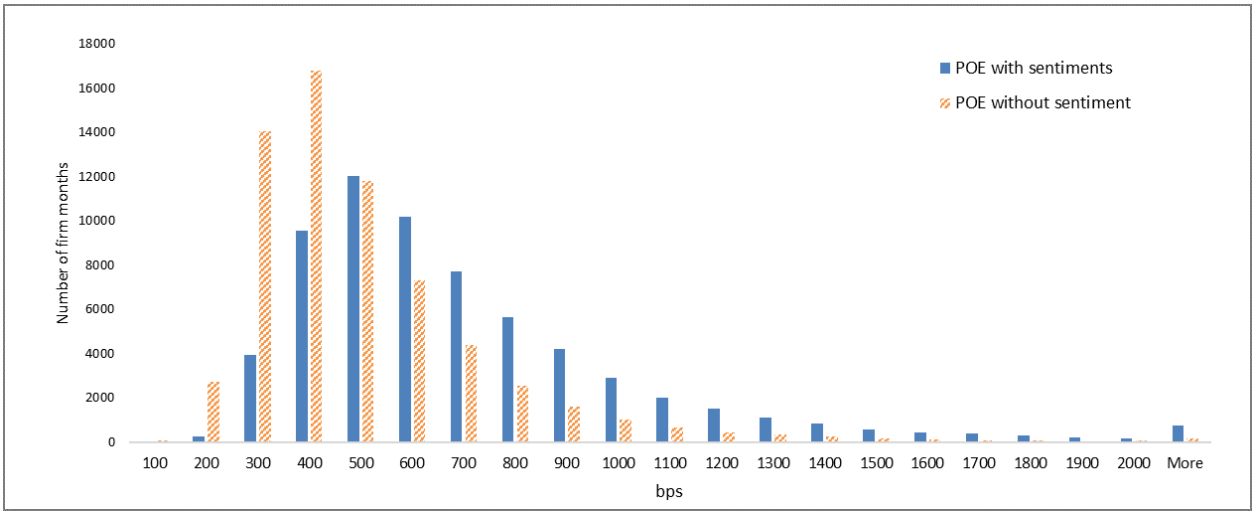


Figure 9: The two POE distributions (with and without media sentiments)

The chance of experiencing other forms of corporate exit, i.e., POE, is also impacted by the inclusion of media sentiments. A glimpse of Figure 9 leads to a conclusion that getting media coverage will increase the likelihood of corporate exit in a form other than default for this sub-sample of firm-month observations. The overall distribution of POEs noticeably shifts to the right and becomes more right-skewed. The earlier statistical results imply that the mere fact of being covered by the three reputed business presses can increase the POE and the impact is more pronounced for those receiving positive sentiments on their credit risks. Recall that other corporate exits are mainly due to mergers and acquisitions. Thus, we can roughly say that media coverage on credit risk is highly suggestive of a firm facing a higher chance of future merger/acquisition, and

in some cases the increase can be very substantial.

Were the probability of merger and acquisition the focal issue, one could further divide the other-exit class into two and use a four-class logistic regression to come up with more focused analysis on mergers and acqusitions.

4.3 Cases to highlight impacts on PD and POE

We take AMC Entertainment Holdings, Inc (AMC) to illustrate concretely the impact on the one-year PD on an individual firm basis and to see how two PD estimates (with and without media sentiments) differ in a time series.

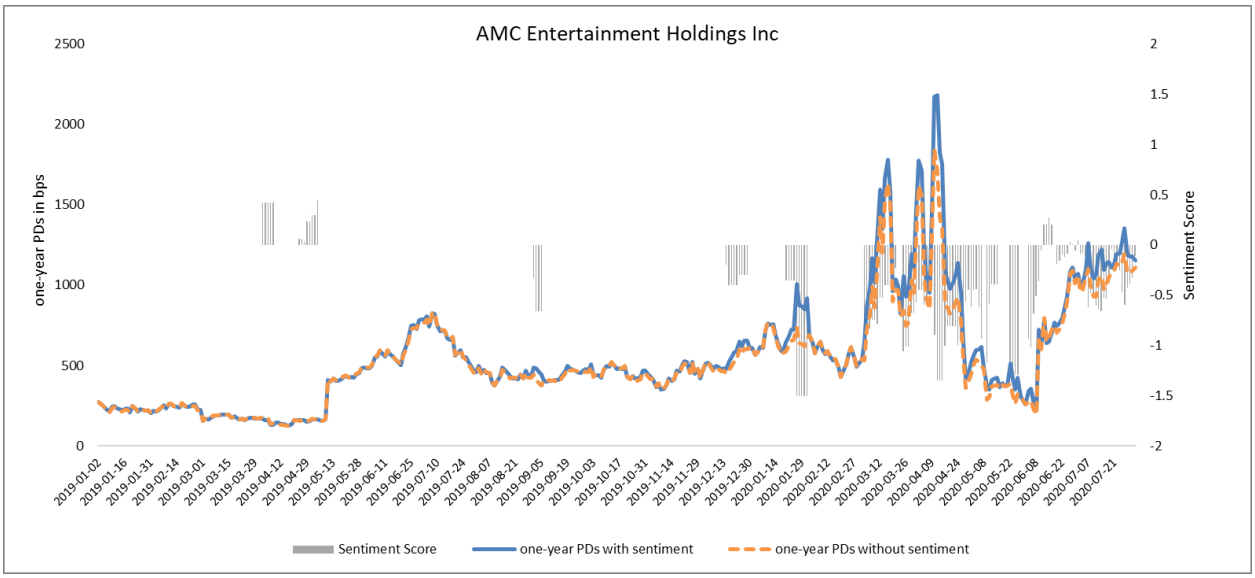


Figure 10: AMC's one-year PDs along with the sentiment scores

AMC's 8k filing dated 31 July 2020 suggests that it entered into a debt exchange agreement with creditors to recognize the losses and to reduce its interest expense burden. Figure 10 shows prior to AMC's default the estimated one-year PDs with the two sentiment variables included and those without. On the same plot, we also display the sentiment scores, which were more negative leading up to the default. The one-year PD with sentiment is higher than that without, and the difference can be quite substantial at times when the sentiments were clearly negative.

Similarly, we use SolarCity Corp ("SolarCity") as an example to show the impact on the estimated chance of experiencing other forms of corporate exit. Figure 11 reports two SolarCity's POE time series, with and without incorporating the two sentiment variables, leading up to the acquisition by Tesla (announced on August 1, 2016, completed on November 21, 2016). Evidently, the POE has significantly shifted upwards throughout the period, responding overall to the mere fact of having a sentiment score and dynamically to a positive sentiment score.

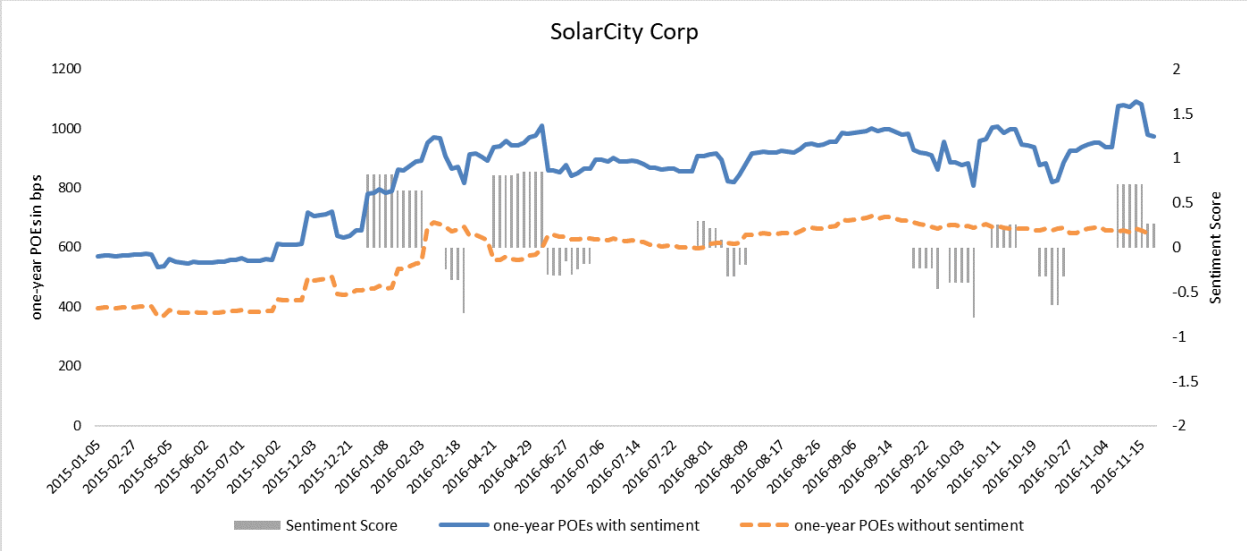


Figure 11: SolarCity's one-year POEs along with the sentiment scores

5 Conclusion

We have put forward an NLP method to produce credit-focused, entity-specific at the article-level sentiments with an aim of complementing already rich structured financial data, aiming to enhance corporate default prediction. We apply the NLP method to a corpus of about 1.7 million articles gathered from three reputable business presses that span 23 years to reach the following conclusions.

First, the credit-focused, entity-specific at the article-level sentiments for those firms with media coverage contain significant additional predictive power above and beyond what the traditional structured financial variables can provide. In addition, the sentiment score along with the sentiment dummy variable can greatly improve the predictive power for other forms of corporate exit.

Although this study focuses on credit risk, the NLP techniques assembled in the paper can be expected to work for other financial and business issues for which one can expect the articles in the business press to offer valuable insight beyond what is already contained in the commonly used structured variables. We can see green finance as one potential area of application.

References

Edward I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4):589–609, 1968.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Samuel W.K. Chan and Mickey W.C. Chong. Sentiment analysis in financial texts. *Decision Support Systems*, 94:53–64, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jin-Chuan Duan and Shuping Li. Enhanced pd-implied ratings by targeting the credit rating migration matrix. *Journal of Finance and Data Science*, 7:115–125, 2021.

Jin-Chuan Duan, Jie Sun, and Tao Wang. Multiperiod corporate default prediction – a forward intensity approach. *Journal of Econometrics*, 170:191–209, 2012.

Darrell Duffie, Leandro Saita, and Ke Wang. Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83:635–665, 2007.

Lee M Dunham and John Garcia. Measuring the effect of investor sentiment on financial distress. *Managerial Finance*, 2021.

Axel Groß-Klußmann, Stephan König, and Markus Ebner. Buzzwords build momentum: Global financial twitter sentiment and the aggregate stock market. *Expert Systems with Applications*, 136:171–186, 2019.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Unpublished software application. https://spacy. io*, 2017.

Xiaodong Li, Pangjing Wu, and Wenpeng Wang. *Information Processing Management*, 57(5): 102212, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Tim Loughran and Bill McDonald. The use of word lists in textual analysis. *Journal of Behavioral Finance*, 16(1):1–11, 2015.

NUS-CRI. *Credit Research Initiative technical report version: 2021 update 1.* 2021a. URL `https://d.nuscri.org/static/pdf/Technical%20report_2021.pdf`.

NUS-CRI. *Probability of Default implied Rating (PDiR2.0).* 2021b. URL `https://d.nuscri.org/static/pdf/PDiR2.0_White_Paper_2021.pdf`.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020.

Jan Roeder, Matthias Palmer, and Jan Muntermann. Utilizing news topics for credit risk management: The explanation of bank cds spreads. *Journal of Decision Systems*, 29(sup1):32–44, 2020.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. *arXiv preprint arXiv:1610.03771*, 2016.

Tyler Shumway. Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business*, 74(1):101–124, 2001.

Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *CoRR*, abs/1903.09588, 2019. URL `http://arxiv.org/abs/1903.09588`.

Yuan Sun, Xuan Liu, Guangyue Chen, Yunhong Hao, and Zuopeng (Justin) Zhang. How mood affects the stock market: Empirical evidence from microblogs. *Information & Management*, 57 (5):103181, 2020.

Feng-Tse Tsai, Hsin-Min Lu, and Mao-Wei Hung. The impact of news articles and corporate disclosure on credit risk valuation. *Journal of Banking and Finance*, 68:100–116, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Andrew M Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal*

Statistical Society: Series B (Methodological), 31(1):80–88, 1969.

Justin Wood. Source-lda: Enhancing probabilistic topic models using prior knowledge sources. *CoRR*, abs/1606.00577, 2016. URL `http://arxiv.org/abs/1606.00577`.

Frank Z. Xing, Erik Cambria, and Yue Zhang. Sentiment-aware volatility forecasting. *Knowledge-Based Systems*, 176:68–76, 2019.

Yang Yu, Wenjing Duan, and Qing Cao. The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4):919–926, 2013.

A Implementing Source-LDA

Implementing Source-LDA requires setting several key hyper parameters. To be specific, we set the number of topics, K, to 2 and the hyper parameters for the Dirichlet prior on the two per-article topic distributions, i.e., α_1 and α_2, to 0.5. The hyper parameters on the prior distribution for the credit risk topic are listed below. Other parameters adopt the default values in Source-LDA. The training process is considered complete after 1,000 iterations.

Hyper parameter values for the prior distribution on the credit risk topic:
bond 10151 market 10000 debt 10000 bank 8727 credit 8000 loan 8000 yield 8000 inflation 8000 default 8000 mature 8000 bond_yield 8000 central_bank 8000 upgrade 8000 downgrade 8000 bankrupt 8000 bankruptcy 8000 bondholder 8000 filing 8000 balance_sheet 8000 cycle 8000 borrowing_cost 8000 financial 6768 collateral 5000 term 5000 corporate 5000 payment 5000 pay 5000 long_term 5000 investment_grade 5000 junk 5000 bps 5000 counterparty 5000 short_term 5000 revenue 5000 gain 5000 profit 5000 earnings 5000 payment 5000 loss 5000 cash 5000 level 5000 banking 5000 fund 4630 risk 4363 growth 4007 economic 3829 high 3829 policy 3295 price 3295 asset 3295 increase 3206 rise 3117 economy 3028 rate 3000 capital 3000 horizon 3000 duration 3000 leverage 3000 liquid 3000 liquidity 3000 income 3000 maturity 3000 pressure 3000 account 3000 finance 3000 company 2671 condition 2582 likely 2493 firm 2404 low 2315 mortgage 2226 demand 2137 cost 2137 credit_risk 2137 sector 2137 issue 2137 investment 2000 investor 2000 government 2000 report 1959 management 1781 swap 1781 note 1781 potential 1692

B TABSA-BERT sentiments

This study adopts a five-category ordinal classification: {-2, -1, 0, 1, 2} representing strongly negative, negative, neutral, positive and strongly positive. The annotation procedure consists of three rounds.

1. The first round: Recruit a finance graduate student to annotate the sentences under guidance.
2. The second round: Invite an individual holding a finance doctoral degree to review the labels and highlight different opinions.
3. The final round: Involve a financial practitioner to discuss those contradictory annotations until a consensus is reached.

We randomly sample 600 articles that are deemed credit risk related (with the credit risk topic weight higher than 0.3). These articles together give rise to 9,208 sentences that have mentioned some companies. These 9,208 sentences form our sample for the TABSA-BERT ordinal classification. The sample is further divided 4 to 1 into training and validation sets. Table A.1 presents the annotated opinion distributions for the training and validation data sets.

Table A.1: Annotation distributions of the fine-tuning dataset

Sentiment	Opinion	# in Training	# in Validation
-2	Strongly Negative	400	136
-1	Negative	1,830	574
0	Neutral	2,908	981
1	Positive	1,477	505
2	Strongly Positive	291	106
Overall		6,906	2,302

For the stochastic gradient descent optimization, we keep the dropout probability at 0.1, set the number of epochs to 20. The best model is selected when the weighted F1 score of the validation set begins to level off. The initial learning rates for BERT and ordinal classification are 1e-5 and 1e-4, respectively. The random seed is set to 42 and the batch size is 16. The fine-tuning performance for the validation data set is presented in Table A.2.

Table A.2: Fine-tuning performance for the validation data set

Sentiment	Opinion	Accuracy	F1 score
-2	Strongly Negative	61.8%	48.4%
-1	Negative	57.3%	62.1%
0	Neutral	82.1%	77.7%
1	Positive	55.8%	60.7%
2	Strongly Positive	42.5%	45.9%
Overall		67.1%	66.9%

Note: The overall Accuracy and F1 score are weighted averages.