BERT應用於預測企業破產

目錄

参考Paper

Alex Kim & Sangwon Yoon *Corporate Bankruptcy Prediction with Domain-Adapted BERT*

Corporate Bankruptcy Prediction with Domain-Adapted BERT

Alex Gunwoo Kim*
Seoul National University
Graduate School of Business
kimgunwoo95@snu.ac.kr

Sangwon Yoon*
Artificial Society Inc.
swyoon@artificial.sc

Abstract

This study performs BERT-based analysis, which is a representative contextualized language model, on corporate disclosure data to predict impending bankruptcies. Prior literature on bankruptcy prediction mainly focuses on developing more sophisticated prediction methodologies with financial variables. However, in our study, we focus on improving the quality of input dataset. Specifically, we employ BERT model to perform sentiment analysis on MD&A disclosures. We show that BERT outperforms dictionary-based predictions and Word2Vec-based predictions under time-discrete logistic hazard model, k-nearest neighbor (kNN-5), and linear kernel support vector machine (SVM). Further, instead of pre-training the BERT model from scratch, we apply self-learning with confidence-based filtering to corporate disclosure data. We achieve the accuracy rate of 91.56% and demonstrate that the domain adaptation procedure brings a significant improvement in prediction accuracy.

corporate disclosures contain non-financial information, textual disclosures have received relatively less attention. Following Li (2008)'s call for research on textual corporate disclosures, there have been numerous attempts (Tetlock et al., 2008; Li, 2010; Mayew et al., 2015) to analyze the textual sentiments of corporate disclosures. They commonly find that textual non-financial information has orthogonal informational value to the existing financial information. However, the majority of the analyses are based on the dictionary-based approach suggested by Loughran and McDonald (2011).

In our study, we perform a BERT-based analysis on corporate disclosure data. BERT (Devlin et al., 2018) is the pre-trained language model based on the self-attention mechanism of Transformers (Vaswani et al., 2017). BERT and its improved versions such as GLUE (Wang et al., 2018), SQUAD (Rajpurkar et al., 2016), and RACE (Lai et al., 2017), have achieved state-of-the-art results in several NLP downstream tasks. In this research, we analyze the management, discussion, and anal-

解決/改善什麼問題

- **重新將資金周轉做分配**：許多企業的破產最主要的原因是資金銜接上出了問題。尤其資金導向型的零售業資金週轉規模驚人，對資金周轉的利用是需要充分考慮，否則會形成長短期資金分配錯誤，將產生潛在的資金鏈斷裂風險。

- **及時瞭解自身行業及供應鏈上下游企業的風險變化**：破產引發的潛在風險可能會給企業信用專家和財務領導者帶來難以估計的壓力，因為他們需要努力應付緩慢的付款，以及處理業務關係，還需要抵禦多米諾骨牌效應的風險。

傳統做法

預測破產

1. 從金融年報中披露的社會變量預測破產(Altman (1968))

2. 與財務報表相關的變量、股票與市場相關的變量 (Shumway(2001))

3. 先前文獻摘錄來自敘述性披露的訊息(Tennyson et al. , 1990)

4. 採用複向量空間模型通過MD&A披露預測破產情況 (Cecchini et al. (2010))

5. 敘述性披露確實包含訊息與提供的訊息正交通過金融變量 (Mayew et al. (2015))

6. 分析MD&A披露的總體基調(Loughran and McDonald (2011))

7. 使用神經網絡與金融預測破產的變量(Wilson and Sharda (1994))

8. 數據包絡分析(DEA) 並表明破產公司表現出相關 (Premachandra et al. (2011))

9. SVM可有效預測顯著企業事件，包括破產(Shin et al. (2005))

10. 開發一個自適應模糊 k-nearest用於破產預測的鄰居方法(Chen et al. (2013))

較少的研究是著重在透過敘述性披露來產生精確的語義語氣分析

文本分類

1. **哈佛心理詞典(Harvard Psychological Dictionary)：**最常用的用於開放域文本分類的源碼，裡面包含正面和負面列表。

2. **金融領域的詞典**：Loughran 和 McDonald (2011) 提出。

基於字典的方法有一個限制，因為字典很難涵蓋所有文本分類需要關鍵字的內容，或是某一些關鍵字出現的頻率不一定有足夠的訊息可以對句子來進行分類。

領域自適應

1.**半監督學習：**當目標域中存在標記數據但數量不足時，使用半監督領域自適應。採用學習帶有標籤的特定任務的距離測量數據並將標籤分配給標記的數據學習到的距離。然而，實際上可能無法找到帶有標記數據的領域。

2.**無監督領域自適應(UDA)：**是一個較好的替代方案。基於子空間的方法將源域和目標域都視為單域空間的子空間。另一方面，無監督領域自適應中更常被使用的方法是考慮源域和目標域分隔空間並嘗試對齊分佈。

如何設計/實現

情感分析

- **基於字典的方法**：Loughran 和 McDonald (2011)開發了適合10-k的檔案的單字表，這個表分別包含了正面和負面的單字，按照計算文本披露語氣的方法，計算每個 MD&A(管理層討論與分析) 部分中正面和負面詞的數量，並按每個部分中的總詞數對其進行縮放。
- **Word2Vec**：使用Word2Vec權重訓練在1996-2003的10-k語料庫，以及使用Malo等人提供的金融情緒分析數據集訓練網路。對文檔中所有句子的概率求和，並將它們歸一化以計算每個文檔（W2VPOS 和 W2VNEG）的情感分數。使用 nltk 句子標記器拆分每個文檔到句子，gensim包加載Word2Vec嵌入，和 Pytorch 實現基於 CNN分類器。使用交叉熵損失函數和亞當優化器。訓練了 60 個 epochs 的模型，批量大小為 50。將句子長度設置為 50訓練階段和推理階段的單詞。

◆**BERT**：利用基於原始 BERT 模型的模型結構和經過微調的 FinBERT 權重進行金融情緒分析。FinBERT 在路透社TRC2 數據集的子集上進行了預訓練，其中包括金融媒體文章，並在金融情緒分析數據集上進行了微調。它與用來訓練 Word2Vec 模型網路的數據集相同。

◆**無監督領域自適應**：首先，生成帶有來自 MD&A 部分句子的標籤。之後過濾掉不相關的樣本句子。因為錯誤的標籤可能會降低性能模型的數量。最後，執行監督使用新獲得的偽標籤進行學習。獲得了 59,389 份與財務變量合併的清理後的 MD&A 文件。

透過隨機挑選 1,200 個文檔並對文檔中的所有句子進行推理來生成偽標籤。具體而言，收集和分析了 589,858 個不同的句子。 之後過濾結果並丟棄觀察結果。透過這個過程，得到了 38,703 個可靠的句子。使用批量大小 32 訓練模型的2個時期，並將學習率設置為 5e−5。 使用交叉熵損失和 Adam 優化器訓練完模型後，推理遵循前兩個模型，產生情緒變量 DAPTPOS 和 DAPTNEG。

模型

比較基線分類器的相對性能，以突出添加基於 BERT 的情感變量的增量預測準確性。

◆**Hazard model**: 計算預測精度

◆**KNN & SVM**: 比較文本情感分類模型

實驗流程

1. 從 Compustat 確定破產公司年數。該數據集提供了公司申請破產的日期和破產程序完成的日期。在分析

中，使用公司首次申請破產的日期作為破產年份。然後，計算了五個已知的先於破產的關鍵財務變量 (WC、RE、EBITDA、MVE 和 SALE_)。

2. 從年度報告中提取 MD&A 部分來構建主要變量。在收集過程中，排除了 html 符號、表格和頁碼。確保

只分析文本MD&A 部分的組件。樣本期從 1995 年到 2020 年。BRUPT是一個指標變量，對於在年度報告發布後 365 天內面臨破產的觀察值等於 1，否則為0。需要每個觀察的所有財務變量和 MD&A 部分文本，並獲得 59,389 個不同的觀察。在樣本中，確定了 520 個破產公司年度觀察結果 (BRUPT=1)。從 Compustat 數據庫中獲取財務數據，並從美國證券交易委員會檔案中獲取歸檔文本。

3. 為了評估 SVM 和 KNN 模型的性能，將樣本分成三個子集：60% 分配給訓練集，20% 分配給驗證集，20% 分配給測試集。選取了 2018-2020 年的 104 個最新的破產觀察，並隨機選取了同一時間的 104 個非破產觀察。為了進一步確保結果不是由隨機樣本選擇驅動的，重複選擇過程 100 次，並報告平均準確度及其標準偏差。
4. 在 SVM 中，測試了 10 個不同的超參數，範圍從 0 到 1，並比較了它們的相對性能。接下來，對於 KNN 模型，實驗了五種不同的超參數 (k)。

結果

主要發現基於BERT的分析優於基於字典的分析和基於 Word2Vec 的分析。這表示特定於上下文的情感分析比非特定於上下文的方法產生更準確的文本基調。具體來說，帶有基於 BERT 的情緒變量的 SVM 顯示了 85.20% 的破產預測準確率。此外，觀察到R平方隨著我們從僅分析金融變量到包括基於域的 BERT 情緒變量而增加。將此結果與先前利用 SVM 預測企業破產的文獻進行比較，獲得了相對較高的準確性。

如何延伸該應用方式

將10-K財報的文本資料透過LDA模型去進行主題分類，觀察出和風險有關的主題下，字詞經由標準化後形成的變數是否能增加破產模型預測的精準度。之後如果要預測一間新公司的財務狀況，則可以直接套用此分類主題的模型去進行預測。

The End