Industry-Sensitive Language Modeling for Business

Philipp Borchert[a,b], Kristof Coussement[a], Jochen De Weerdt[b], Arno De Caigny[a]

[a] *IESEG School of Management, Univ. Lille, CNRS, UMR 9221 - LEM - Lille Economie Management, 3 Rue de la Digue, Lille, F-59000, France*
[b] *KU Leuven, Naamsestraat 69, Leuven, 3000, Belgium*

Abstract

We introduce BusinessBERT, a new industry-sensitive language model for business applications. The key novelty of our model lies in incorporating industry information to enhance decision-making in business-related natural language processing (NLP) tasks. BusinessBERT extends the Bidirectional Encoder Representations from Transformers (BERT) architecture by embedding industry information during pretraining through two innovative approaches that enable BusinessBert to capture industry-specific terminology: (1) BusinessBERT is trained on business communication corpora totaling 2.23 billion tokens consisting of company website content, MD&A statements and scientific papers in the business domain; (2) we employ industry classification as an additional pretraining objective. Our results suggest that BusinessBERT improves data-driven decision-making by providing superior performance on business-related NLP tasks. Our experiments cover 7 benchmark datasets that include text classification, named entity recognition, sentiment analysis, and question-answering tasks. Additionally, this paper reduces the complexity of using BusinessBERT for other NLP applications by making it freely available as a pretrained language model to the business community. The model, its pretraining corpora and corresponding code snippets are accessible via `https://github.com/pnborchert/BusinessBERT`.

Keywords: Analytics, Natural Language Processing, OR in business, Artificial Intelligence

1. Introduction

The availability and importance of textual data in analytical models that support business decision making is increasing (Frankel et al., 2021). This is reflected in the growing number of studies

Email addresses: `p.borchert@ieseg.fr`, `philipp.borchert@kuleuven.be` (Philipp Borchert), `k.coussement@ieseg.fr` (Kristof Coussement), `jochen.deweerdt@kuleuven.be` (Jochen De Weerdt), `a.de-caigny@ieseg.fr` (Arno De Caigny)

in the business literature that highlight the added value of analyzing textual data, for example, in applications such as financial risk prediction (Bao and Datta, 2014; Stevenson et al., 2021), corporate innovation prediction (Bellstam et al., 2021) or news-based macroeconomic forecasting (Feuerriegel and Gordon, 2019). In fact, this increased adoption of natural language processing (NLP) can be observed in a variety of business disciplines, as shown in **Table 1**. Modern approaches to analyzing textual data rely on transfer learning, which includes pretraining language models on large amounts of textual data, which is hardly related to the downstream task. Therefore, the pretrained language model is further fine-tuned on task-specific data. Yet, a crucial challenge arises when general-purpose language models are applied in a business context. The general-purpose corpora used to train the language model mismatch the industry-specific vocabulary and terminology used in downstream applications. In turn, this causes a decrease in contextual relevance of the textual representations, which negatively impacts model performance. Hence, we propose BusinessBERT, a language model trained on large-scale business communication corpora to create industry-sensitive vector representations of textual data.

In an influential study, Devlin et al. (2019) proposed BERT - Bidirectional Encoder Representations from Transformers - as a general-purpose language model. As a transformer encoder, BERT creates contextualized vector representations of textual data. The model is pretrained on general-purpose corpora such as English Wikipedia and BooksCorpus through predicting masked words. This extensive pretraining process exposes the model to a variety of natural language. In transfer learning, this information is leveraged to increase performance and reduce the amount of labeled data required to fine-tune the model for specific downstream tasks.

Recent challenges for researchers and practitioners incorporating textual data in business applications include resource intensive training, limited access to task-specific text, and manual data labeling. Furthermore, the application of dictionary approaches restricts the use of external information and knowledge in addition to the task-specific text (Archak et al., 2011; Lee et al., 2018; Zhang and Luo, 2022; Xu et al., 2022). As a business-specific language model, BusinessBERT addresses these challenges and reduces the complexity of leveraging textual data in operational research (OR) applications.

Several BERT extensions that display outstanding performance in various domains and downstream tasks have been developed (see, e.g., Beltagy et al. (2019), Lee et al. (2019) and Araci

Table 1: Overview business-related NLP applications by discipline

Discipline	Applications
Business Strategy	negotiation communication analysis (Jeong et al., 2019), competitor analysis (Pan et al., 2019; Liu et al., 2020), news-based network analysis (Chen et al., 2021a), corporate diversification analysis (Choi et al., 2021)
Finance	financial risk classification (Bao and Datta, 2014), news-based macroeconomic forecasting (Feuerriegel and Gordon, 2019), business failure prediction (Stevenson et al., 2021; Wang et al., 2021a; Borchert et al., 2022), risk perception prediction (Bhatia, 2019), loan default prediction (Netzer et al., 2019), corporate risk analysis (Hsu et al., 2022), financial forecasting (Díaz et al., 2023), merger prediction (Katsafados et al., 2024)
Innovation Management	patent similarity analysis (Arts et al., 2018), corporate innovation prediction (Bellstam et al., 2021), patent classification (Miric et al., 2022)
Marketing	social media engagement prediction (Lee et al., 2018; Li and Xie, 2020), extracting marketing information from social media (Hartmann et al., 2021), customer service perception analysis (Puranam et al., 2021), learning engagement analysis (Narang et al., 2022), marketing appeal generation (Hong and Hoban, 2022) customer complaint analysis (Vairetti et al., 2024), customer experience (Aldunate et al., 2022)
Operations	marketplace reputation analysis (Moreno and Terwiesch, 2014), social media engagement prediction (Lee et al., 2018; Shin et al., 2020), hospital readmission prediction (Baechle et al., 2020), health-related question answering (Mousavi et al., 2020), service quality analysis (Xu et al., 2021), employee review analysis (Symitsi et al., 2021), review helpfulness prediction (Liu et al., 2021)

(2019); Yang et al. (2020)). However, existing language models are suboptimal for a variety of business applications, due to a mismatch of pretraining on general-purpose language corpora and industry-specific textual data in downstream applications.

Thus, we introduce BusinessBERT, a BERT-based language model trained on business communication corpora that innovates on the pretraining objectives with the goal of capturing industry information in textual data. The key characteristics of BusinessBERT include training the model on large-scale business communication corpora and using industry classification as a pretraining objective to create industry-sensitive representations of textual data. First, we collected business-relevant textual content from company websites, management discussion & analysis (MD&A) disclosures and scientific papers in the business domain as input to BusinessBERT. These corpora contain 2.23 billion tokens for training BusinessBERT, are publicly accessible, and cover a wide variety of business communication topics in different industries.

Second, we introduce industry classification (IC) as a pretraining objective to develop industry-sensitive language representations for business-related NLP applications. This process involves assigning an industry category based on standard industry codes (SICs) to each textual input during pretraining. By including IC during pretraining, the textual data representations retain industry information. The use of IC as a pretraining objective is inspired by recent studies finding company-related textual content and terminology to differ in line with industry categories (Hoberg and Phillips, 2016; Shi et al., 2016; Xu et al., 2020). We show that capturing information according to different industry categories supports contextual understanding of business-related text and therefore improves performance on downstream tasks. **Figure 1** displays an example of the IC objective performed by BusinessBERT during pretraining.

To demonstrate the performance of BusinessBERT, we investigate four NLP tasks that can be considered the building blocks of a large majority of downstream NLP applications in the business and OR literature (Bao and Datta, 2014; Frankel et al., 2021; Deng et al., 2018). NLP tasks include text classification (Bao and Datta, 2014), named entity recognition (Geng et al., 2021), sentiment analysis (Frankel et al., 2021; Deng et al., 2018) and question-answering (Chen et al., 2021b). A mapping of selected NLP business applications in **Table 1**) to the four NLP tasks is provided in **Appendix A2**.

We compare the performance of BusinessBERT compared to state-of-the-art language models

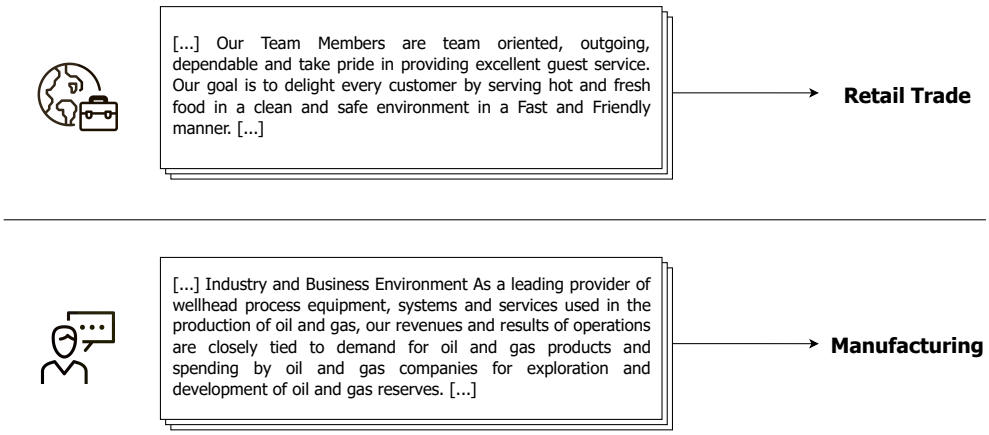

Figure 1: Industry classification example

based on overlapping model architectures and target domains, namely BERT-Base (Devlin et al., 2019) and FinBERT (Araci, 2019; Yang et al., 2020). In general, we find that BusinessBERT benefits from industry-sensitive training using business communication corpora and industry classification as an objective, demonstrating substantial performance improvements in three of the four business-related NLP tasks, with average performance improvements between 1.00% and 6.39% over BERT-Base, between 2.14% and 9.69% over FinBERT (Araci, 2019) and between 0.71% and 28.08% over FinBERT (Yang et al., 2020). Furthermore, our training approach requires significantly less pretraining data than other benchmark models, ranging between -28% and -54%. We validate the transferability of our findings to other language models, by fine-tuning RoBERTa (Liu et al., 2019) and LLaMA 2 (Touvron et al., 2023) on our business communication corpora using the IC objective. The results indicate that RoBERTa benefits from industry-sensitive fine-tuning on specific datasets, while industry-sensitive fine-tuning improves LLaMA 2 performance for all business-related NLP tasks.

BusinessBERT is freely available for researchers and practitioners and can be easily integrated into downstream applications using the code snippets provided [1]. Furthermore, the CompanyWeb [2] corpus, which consists of textual content extracted from more than 1.7 million web pages of more than 390,000 companies is also available.

[1]The pretrained PyTorch checkpoint is available at `https://huggingface.co/pborchert/BusinessBERT`. The corresponding GitHub repository and code snippets are available at `https://github.com/pnborchert/BusinessBERT`.

[2]The dataset, including the SIC codes, is available at `https://huggingface.co/datasets/pborchert/CompanyWeb`.

2. BusinessBERT

We propose BusinessBERT as a pretrained language model for NLP applications in the business domain. BusinessBERT is pretrained on business communication extracted from company websites, management discussion and analysis (MD&A) statements and the business section of the Semantic Scholar Open Research Corpus (S2ORC). By annotating the corpus with standard industry classification (SIC) labels, BusinessBERT performs industry classification (IC), as well as masked language modeling (MLM) and next sentence prediction (NSP) as pretraining objectives. The training approach, including the corpora and pretraining objectives of BusinessBERT is summarized in **Figure 2**.

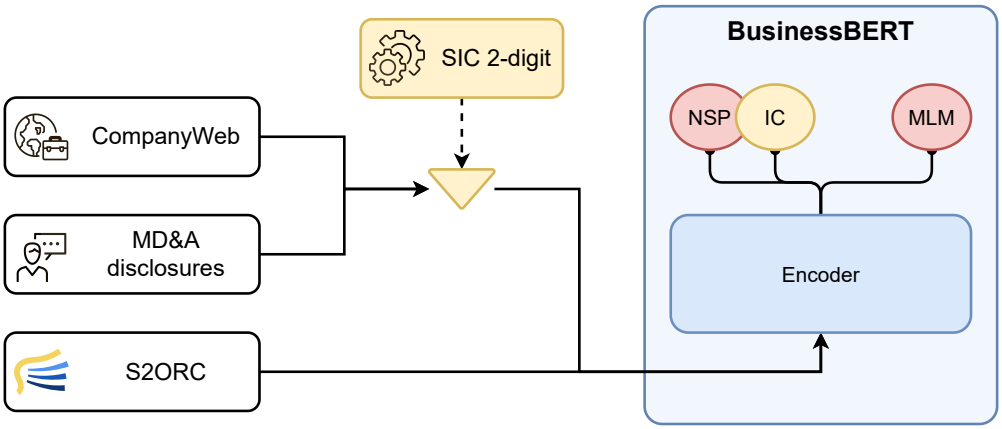

Figure 2: BusinessBERT training approach

2.1. BERT

Recent advances in NLP demonstrate the performance of self-supervised transformer models (Devlin et al., 2019). Transformer models consist of a stacked encoder-decoder structure, with the encoder creating fixed-length vector representations of the textual input and the decoder generating text sequences with various lengths as output. The encoder and decoder both incorporate multi-head attention, which enables the model to capture multiple long-term dependencies in sequences (Vaswani et al., 2017).

Vaswani et al. (2017) utilize scaled dot product attention, which computes the weighted sum of the values V defined by query Q for each corresponding key K. The term self-attention refers to the case where the query, keys, and values originate from the same sequence.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Scaling the dot product by the square root of the key dimension d_k prevents small gradients from prematurely stopping weight updates during the training process.

Multi-head attention allows the model to capture multiple dependencies in one sequence by concatenating the results of multiple attention transformations performed in parallel. Multi-head attention is defined as follows:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \tag{2}$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ and the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d\times d_k}$, $W_i^K \in \mathbb{R}^{d\times d_k}$, $W_i^V \in \mathbb{R}^{d\times d_v}$ and $W_i^O \in \mathbb{R}^{hd_v\times d}$ with h attention heads and d projection dimensions (Vaswani et al., 2017).

Before input of textual content into the model, the text is encoded using key-value pairs of (sub) words and their categorical values (tokens). In addition to the token value, the positional information of the token in the sequence is retained by positional encodings (Vaswani et al., 2017). The tokens are projected onto d dimensional vectors by an embedding layer before they are processed in the encoder-decoder blocks. An encoder (or decoder) block refers to architecturally identical layers that can be stacked sequentially. Encoder blocks are composed of multi-head self-attention, feedforward layers, and residual connections, followed by layer normalization. The encoder architecture is visualized in **Figure 3**. The decoder block extends the encoder block, employing an additional multi-head attention layer over the output of the encoder blocks (Vaswani et al., 2017).

Devlin et al. (2019) developed BERT as a general-purpose language model based on the encoder described by Vaswani et al. (2017). The BERT-Base model includes $N = 12$ encoder blocks and $h = 12$ self-attention heads with $d = 768$ dimensional vector representations (hidden size). The model is trained on English text originating from the English Wikipedia and BooksCorpus (3.1 billion tokens) using self-supervision. This resource-intensive pretraining phase reduces the amount of time and labeled data needed to fine-tune BERT to perform specific tasks. Due to the encoder architecture, the model creates contextualized vector representations of textual input data that can be used in any downstream task, such as text classification or question-answering. BERT utilizes a vocabulary of 30,000 subwords and WordPiece encodings to represent textual inputs as tokens.

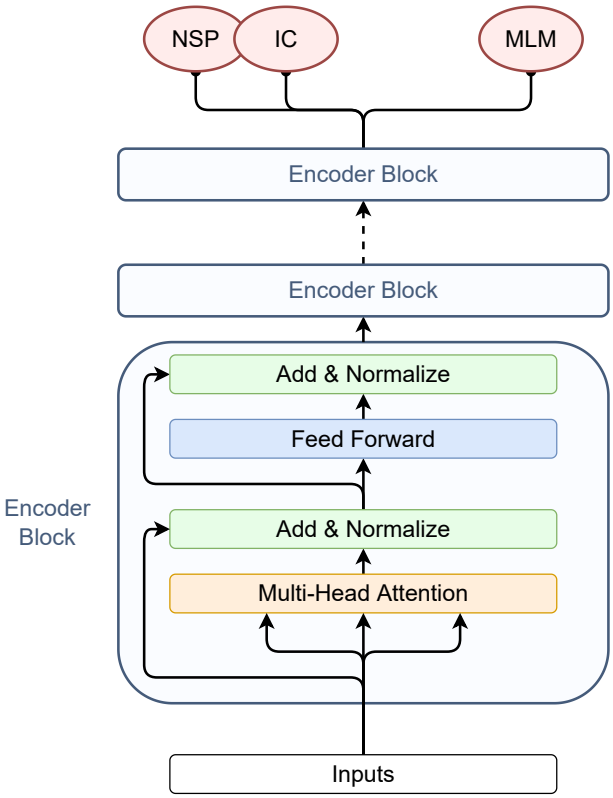

Figure 3: Encoder architecture including the IC objective (Vaswani et al., 2017; Devlin et al., 2019).

Devlin et al. (2019) defined two tasks that are performed during self-supervised pretraining: masked language modeling (MLM) and next sentence prediction (NSP). First, during MLM, 15% of the provided input tokens are either randomly replaced with a "[MASK]", a random token in the text corpus, or are left unchanged. The masked tokens are predicted based on all nonmasked tokens in the sequence, thereby exploiting bidirectional context. Moreover, the objective of the NSP task is to improve the model's performance in downstream NLP tasks, such as question-answering (Devlin et al., 2019). To encourage BERT to incorporate contextual dependencies across sentences, the inputs are processed in sentence pairs during pretraining. Furthermore, the "[CLS]" token is appended to the inputs and subsequently used as the sentence embedding. The model performs binary classification, predicting whether the input consists of subsequent sentence pairs. Thus, sentences are randomly sampled with a 50% probability of being either subsequent sentences or sentences from a different document within the corpus.

2.2. Corpora & Vocabulary

The dataset used to train BusinessBERT includes three large-scale corpora that represent popular sources of business communication (Lo et al., 2020; Ewens, 2019). We ensured that the dataset contained high-quality textual data sources, by removing duplicates, as well as documents with less than 150 tokens. Additionally, we ensure a minimum sequence length of 10 tokens. **Table 2** summarizes the corpora used in the BusinessBERT pretraining process.

Table 2: Overview corpora

Corpus	#Tokens (in billion)	#Documents	Avg. #tokens per document	Size (in GB)[1]	Source
CompanyWeb	0.77	1,788,413	281.89	3.5	**This study**
MD&A	1.06	2,393,595	442.85	5.1	Ewens (2019)
S2ORC	0.40	1,723,517	232.08	1.9	Lo et al. (2020)

[1] Size of the uncompressed text file in GB.

Company websites are inherently rich data sources that contain textual information directed toward various company stakeholders including customers, suppliers, investors and employees (Borchert et al., 2022). The textual content on a company website is a self-reported digital representation of the company, that contains terminology regarding marketing, finance, customer relationships and industry specific terms. Recent business applications that incorporate textual web page data include assessing the projected employer brand image (Theurer et al., 2022). We gather textual content from more than two million web pages on 393,542 company websites. The list of companies was compiled using the Orbis database of Bureau Van Dijk (Bureau van Dijk, 2021), which we enriched with general company information such as standard industry classification (SIC) labels. The dataset includes small, medium and large international enterprises including publicly listed companies. We extracted all available textual information starting from the website homepage between 2014 and 2021. To ensure the extraction of relevant information, our search included all linked subsequent pages accessible from the homepage that contain the company domain name. Additionally, we removed pop-ups and cookie banners during the extraction of website content. We filtered the resulting textual data to include only English text using the FastText language detection API (Joulin et al., 2016). Further preprocessing steps include the removal of duplicate sites and pages with overlapping textual content. The preprocessed CompanyWeb corpus includes

1,788,413 text documents containing 0.77 billion tokens.

Qualitative data included in periodic disclosures are regarded valuable sources of information to evaluate management opinions and forecasts and complement quantitative data in business analysis (Beyer et al., 2010; Bao and Datta, 2014; Feldman et al., 2010; Wang et al., 2021b; Frankel et al., 2021). The MD&A section is part of the periodic disclosures to the Securities and Exchange Commission (SEC) and is therefore available for publicly listed companies in the US. Due to their wide availability and close relation to financial ratios, MD&A statements are commonly used as textual data sources in financial applications (Purda and Skillicorn, 2015; Wang, 2021). MD&A disclosures include textual content regarding a firms' risk assessment, future goals and competition analyses. The statement is written by management and must depict a balanced view of the current and future position of the company. Moreover, these disclosures are directed directly at analysts and investors (Li, 2010; Wang, 2021). We used the textual content of MD&A statements published by 16,130 companies between 2002 and 2018 (Ewens, 2019). The dataset was enriched with firmographic data, including SIC labels.

The Semantic Scholar Open Research Corpus (S2ORC) contains a wide range of scientific research papers with medicine, biology and chemistry as the most prominent categories (Lo et al., 2020). We selected 1.8 million abstracts and 94,000 full-text papers that indicated "Business" as the field of study. The corpus includes scientific papers written before May 2020. Academic research papers do not serve as direct communication media for addressing company stakeholders and, therefore, contain an external perspective on business communication while increasing further language variety and introducing novel terminology. Beltagy et al. (2019) previously employed the entire S2ORC as a valuable data source in the development of SciBERT.

Table 3: Lexical overlap between pretraining corpora (in %).

	CompanyWeb	MD&A	S2ORC
CompanyWeb	100.00	45.57	45.93
MD&A	45.57	100.00	44.86
S2ORC	45.93	44.86	100.00

In Table 3, we assess the lexical diversity across the business communication corpora by examining the overlap between the 10,000 most frequent words in each of the three corpora (Gururangan

et al., 2020). We observe lexical overlap ranging from 44.86% to 45.93%, indicating significant lexical diversity among these corpora, despite their shared domain. In accordance with Lee et al. (2019); Yang et al. (2020), we developed a domain-specific vocabulary based on the three business communication corpora. We established the BusinessBERT vocabulary using WordPiece encodings (Devlin et al., 2019) and the tensor2tensor library (Vaswani et al., 2018). Careful adjustments were made to the vocabulary, excluding characters from non-English languages. This not only ensured an efficient training process but also prevented the use of unknown tokens. As a result, the vocabulary comprises 29,389 tokens. In line with (Devlin et al., 2019), we included 1,000 unused tokens in the vocabulary to provide the possibility of incorporating specialized terminology.

2.3. Industry Classification

We introduce IC derived from two-digit SICs as a new pretraining objective for language models. IC aims to establish a vocabulary for industry-specific terminology while composing the model's understanding of business communication. IC codes such as SIC are commonly used to compare companies among corporate peers and control for differences between industry segments (Davis et al., 2012; Koo et al., 2017; Nauhaus et al., 2021; Sun et al., 2021). Frankel et al. (2021) transferred this concept to textual data by including fixed industry effects based on textual sentiment features in their analysis. Hoberg and Phillips (2016) introduced a text-based industry classification model that used business descriptions in 10-K filings. These studies provide strong evidence that the differences between industries permeate textual information such as MD&A statements or corporate website content.

In the following section, we describe the implementation of the new IC objective in the BusinessBERT pretraining process. Given a text sequence x, we utilize the sequence representation [CLS] to predict the corresponding industry code $i \in \mathcal{I}$, with the probability of predicting industry i denoted as:

$$p(i \mid x) = \frac{\exp\left(w_i \cdot h_{[\mathrm{CLS}]}\right)}{\sum_{i' \in \mathcal{I}} \exp\left(w_{i'} \cdot h_{[\mathrm{CLS}]}\right)}, \tag{3}$$

where $h_{[\mathrm{CLS}]}$ is the hidden vector of [CLS] and w_i denotes the pre-softmax vector corresponding to $i \in \mathcal{I}$. We consider the two-digit SICs for the categories displayed in **Table 4**, including an additional category to represent the absence of industry classification data ("NA"). As a result, the IC objective focuses on differentiating industry affiliation at the least granular aggregation level. In line with Hoberg and Phillips (2016), we argue that with increasing SIC granularity,

Table 4: Standard Industry Classification Codes

Code	Industry
01-09	Agriculture, Forestry, Fishing
10-14	Mining
15-17	Construction
20-39	Manufacturing
40-49	Transportation, Communications, Electric, Gas, Sanitary Services
50-51	Wholesale Trade
52-59	Retail Trade
60-67	Finance, Insurance, Real Estate
70-89	Services
90-99	Public Administration

company affiliations with individual categories become more ambiguous. Text documents in the CompanyWeb and MD&A corpora were acquired from companies and are annotated with the SIC labels of the respective firm. In instances where companies have multiple SIC codes, we opt for the primary code to retain one target value for each input, resulting in a multi-class classification task. Notably, approximately 33.81% of the companies within our sample incorporate secondary SIC labels, with only 11.18% diverging from the primary industry. Therefore, redefining the IC as a multi-class multi-label objective would introduce training inefficiencies due to additional computational complexity. The distribution of SIC labels in the corpora, displayed in **Figure 4**, corresponds to the SIC label distribution in financial databases, such as Orbis and Compustat (Liu et al., 2020). Text documents in the S2ORC corpus do not correspond to specific industries and are therefore not annotated with SIC codes. During pretraining we compute the total loss as the sum of the individual MLM, NSP and IC loss terms for the CompanyWeb and MD&A corpora while masking the IC loss term for the S2ORC corpus.

To validate the contribution of the IC loss to the total loss and therefore its relevance during pretraining, we display the model performance of the three pretraining objectives on a randomly sampled holdout set of corpora in **Figure 5**. We find the accuracy of IC and MLM to increase over the training steps. With higher accuracy on the IC compared to the MLM objective, IC contributes

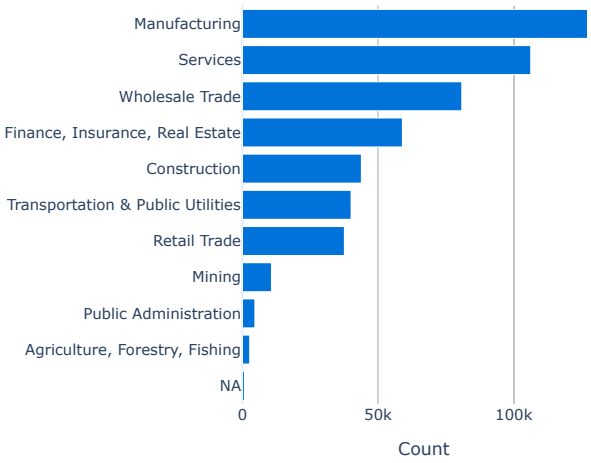

Figure 4: SIC Distribution

significantly to the total loss during pretraining, while remaining below (almost) perfect prediction accuracy. While the contribution to the loss during pretraining validates the relevance of the IC to model convergence, it does not evaluate the performance of the model in downstream applications, which is included in **Section 4**.

2.4. Architecture & Training

In line with Araci (2019) and Yang et al. (2020), we select the BERT-Base model architecture, i.e., $N = 12$ encoder blocks and $h = 12$ self-attention heads with $d = 768$ dimensional vector representations (hidden size) and a maximum input size of 512 tokens. BusinessBERT was trained with a batch size of 128 for approximately 70 epochs using the Adam optimizer with a linearly decaying learning rate of 5e-5 (Devlin et al., 2019). The model is trained for approximately 96 hours on v2-TPUs provided by Google's TPU Research Cloud.

3. Empirical Evidence

3.1. Data

We evaluate the performance of BusinessBERT on four NLP tasks including text classification, named entity recognition, sentiment analysis, and question-answering, which commonly serve as a

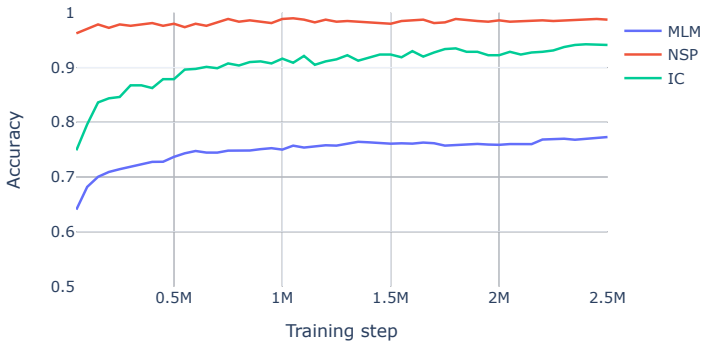

Figure 5: Accuracy pretraining objectives

building block for popular business applications of NLP (see **Table 1**). The benchmark datasets, all from the business domain, are openly accessible to facilitate reproducible results and straightforward transfer to proprietary data. In the following section, we describe the benchmarks used to evaluate the performance of BusinessBERT. An overview of the benchmark datasets is provided in **Table 5**.

In *text classification* tasks, a model presents text sequences as context and infers the corresponding class. Various business-related NLP applications can be expressed as a text classification task, including patent classification (Miric et al., 2022), financial distress prediction (Wang et al., 2021b), social media engagement prediction (Shin et al., 2020) and hospital readmission prediction (Baechle et al., 2020). The benchmark includes text classification applications for risk prediction and news topic classification:

- Bao and Datta (2014) extracted textual information from the 1A section in the 10-K disclosures of 122 companies and used this data to enhance risk classification. Text sequences are selected to identify one of 27 types of financial risk. The authors reported the classification accuracy of the Sent-LDA-VEM (82.55%) and CKNN (83.62%) models based on five-fold cross-validation (Bao and Datta, 2014).

- The Reuters collection contains 21,578 news headlines and the corresponding topic labels (Hayes, 1992). We utilize this news dataset to evaluate model performance in distilling embedded topics

Table 5: Overview of datasets from the business domain per NLP task

NLP Task	Name	Textual data	Avg. # tokens	Target	Source
Text classification	Risk	1A section included in 10-K disclosures	27.86	Risk type {1, . . . , 27}	Bao and Datta (2014)
	News	News headlines	159.83	News topic {1, . . . , 55}	Hayes (1992)
Named entity recognition	SEC filings	Financial agreements (SEC filings)	37.23	Entity type {LOC, ORG, PER, MISC}	Alvarado et al. (2015)
Sentiment analysis	FiQA	Microblog messages, news statements and headlines	22.77	Sentiment score [-1, 1]	Maia et al. (2018)
	Financial Phrasebank	Financial news articles	29.10	Sentiment type {positive, neutral, negative}	Malo et al. (2014)
	StockTweets	Twitter messages	51.27	Sentiment type {positive, neutral, negative}	Taborda et al. (2021)
Question answering	FinQA	Coporate earnings reports	146.61	Reasoning program $\{w_0, w_1, \ldots, w_n\}$[1]	Chen et al. (2021b)

[1] w_i are program tokens that represent mathematical operations.

from (short) text sequences. News headlines can be labeled with multiple topics; thus, this problem is a multilabel classification task. We filter the dataset to include topics that appear at least 20 times, retaining 55 distinct topics.

Named entity recognition tasks involve extracting the entity type corresponding to individual tokens or words in text documents. It is commonly employed in automated document processing and can be used to enhance relation extraction (Geng et al., 2021), news-based network analysis (Chen et al., 2021a), part-of-speech tagging (Xu et al., 2021; Li and Xie, 2020) or document retrieval systems (Li et al., 2021).

- The SEC filings dataset contains textual content extracted from financial agreements included in corporate SEC filings annotated in the CoNLL-2003 format (Alvarado et al., 2015). The dataset includes location (LOC), organization (ORG), person (PER) and miscellaneous (MISC) entities annotated in the text documents.

Sentiment analysis is commonly used to quantify textual data by extracting opinions or tones embedded in written text (Wang et al., 2021b; Deng et al., 2018), i.e. Wang et al. (2021b) extracted sentiment information from annual reports to enhance the performance of default prediction and financial distress prediction models and Puranam et al. (2021) utilized BERT to analyze consumer sentiment in online reviews. Business-related NLP applications leverage sentiment information to enrich downstream applications with textual features and enhance predictive performance, however, sentiment analysis is not used as a downstream application itself. Sentiment analysis tasks can be expressed as classification or regression problems by defining prediction targets according to sentiment categories or continuous sentiment scores.

- The first part of the FiQA dataset includes finance-related microblog messages, news statements or headlines labeled with continuous sentiment scores from "negative" (-1) to "positive" (1). We do not utilize the aspect annotation included in the dataset to ensure that the sentiment analysis tasks can be compared across datasets.

- The Financial PhraseBank introduced by (Malo et al., 2014) contains text sequences from financial news articles annotated with the sentiment categories "positive", "neutral", and "negative". In our evaluation, we select 2264 samples with 100 percent agreement between annotators.

- StockTweets contains 1,300 tweets published in 2020 using tags such as "#SPX500" and "#stocks". The text is annotated with a categorical sentiment label as "positive", "neutral" or "negative" (Taborda et al., 2021).

Question answering involves expressing the tasks as natural language questions that are used as textual input to the model. Depending on the task, the answer can include, i.e. text classification, sentiment analysis or the generation of text sequences. With the flexibility provided by formulating tasks in question-answer format, various business-related NLP applications including all aforementioned tasks can be expressed using question answering. Recent applications of question-answering in business literature include information extraction and retrieval (Mousavi et al., 2020; Chen et al., 2021b) and generative tasks (Chen et al., 2021b), where the question and the additional textual context are provided as input to the model.

- The FinQA dataset includes earnings reports of S&P 500 companies between 1999 to 2019 and contains textual context in the form of documents and tables (Chen et al., 2021b). Questions and reasoning programs that contain the mathematical derivations required to obtain the correct answer are provided as inputs for the model. The generated answers are evaluated according to the solution and numerical reasoning using mathematical operators (i.e. add, divide). We follow the retriever-generator architecture proposed by Chen et al. (2021b), using BERT-Base to retrieve the most relevant documents for each question. This process ensures that the data available for subsequent answer generation are equivalent for all benchmark models. The generator consists of the benchmarked transformer (encoder) and a long short-term memory (LSTM) network (decoder) (Chen et al., 2021b).

In line with the analysis in **Section 2.2**, we evaluate the lexical diversity among benchmark datasets by comparing the overlap between their most frequent words (Gururangan et al., 2020). As shown in **Table 6**, the FinQA and News datasets display the largest lexical overlap at 35.05%. Conversely, the overlap among all other datasets remains below 30%. In summary, the benchmark datasets provide a lexically diverse benchmark for business-related NLP tasks.

3.2. Model

We evaluate BusinessBERT according to two benchmark models, namely BERT-Base (Devlin et al., 2019) and FinBERT (Araci, 2019; Yang et al., 2020), based on their similar model architecture

Table 6: Lexical overlap between benchmark datasets (in %).

	Risk	News	SEC filings	FiQA	Fin. Phrasebank	StockTweets	FinQA
Risk	100.00	19.77	27.06	13.37	18.89	11.63	22.35
News	19.77	100.00	18.58	15.67	24.84	17.05	35.05
SEC filings	27.06	18.58	100.00	11.95	17.45	10.45	20.70
FiQA	13.37	15.67	11.95	100.00	15.79	18.32	13.04
Fin. Phrasebank	18.89	24.84	17.45	15.79	100.00	14.85	22.63
StockTweets	11.63	17.05	10.45	18.32	14.85	100.00	14.90
FinQA	22.35	35.05	20.70	13.04	22.63	14.90	100.00

and target domains. We provide an overview of the benchmark models, including their respective input corpora and pretraining objectives in **Table 7**.

Table 7: Overview models

Model	Architecture	#Parameters (in million)	Corpora	#Token (in billion)	Training objectives
BERT-Base	BERT-Base	110	BooksCorpus English Wikipedia	3.10	MLM, NSP
FinBERT (Araci, 2019)	BERT-Base	110	BooksCorpus English Wikipedia TRC2	3.13	MLM, NSP
FinBERT (Yang et al., 2020)	BERT-Base	110	Corporate Reports Earnings Call Transcripts Analyst Reports	4.90	MLM, NSP
BusinessBERT	BERT-Base	110	CompanyWeb MD&A Disclosures S2ORC	2.23	MLM, NSP, IC

The set of hyperparameters used to fine-tune the models is described in **Table 8** (Devlin et al., 2019). The reported model performance is based on 10-fold cross-validation. The AdamW optimizer was used to fine-tune the models (Loshchilov and Hutter, 2019). The outputs were obtained from a single output layer subsequent to the transformer model. In line with the model suggested by Chen et al. (2021b), we employ an LSTM decoder for the FinQA dataset. To facilitate

ease of implementation and reproduceability of our results, the models are fine-tuned using the transformers library (Wolf et al., 2020) with code snippets available through our GitHub repository. The experiments are conducted on a NVIDIA RTX 4000 GPU.

Table 8: Fine-tuning hyperparameters

Hyperparameter	Value
Learning Rate	{1e-5, 2e-5, 3e-5, 4e-5, 5e-5}
Batch Size	{16, 32}
Max. Epochs	20
Warmup Proportion	0.1
Learning Rate Decay	Linear
Regularization	{None, Early Stopping}

4. Results

Compared to BERT-Base and FinBERT, BusinessBERT shows substantially better performance in the text classification, named entity recognition and question-answering tasks. We present the results aggregated by downstream task in **Tables 9,10,11** and **12**. The tables include the performance values averaged over ten cross-validation folds with the respective standard deviation values.

4.1. Text Classification

For the risk and news classification datasets, BusinessBERT outperforms the other benchmark models with an average performance improvement of 3.90% over BERT-Base. BERT-Base and FinBERT (Araci, 2019) show the worst performance among the benchmark models. With the exception of FinBERT (Araci, 2019), all benchmarked models outperform the CKNN baseline (83.62% accuracy (Bao and Datta, 2014)) in the risk dataset. The domain-specific FinBERT (Yang et al., 2020) shows significantly better performance in comparison with BERT-Base.

4.2. Named Entity Recognition

Based on the results in **Table 10** BusinessBERT outperforms all benchmark models in the named entity recognition task, showing a 1.24% performance improvement over BERT-Base. More-

Table 9: Results: Text classification

	Risk		News	
	F1	Acc	F1	Acc
BERT-Base	82.89 ±0.01	84.58 ±0.02	71.41 ±0.02	65.12 ±0.02
FinBERT (Araci, 2019)	81.35 ±0.02	83.28 ±0.02	68.81 ±0.04	63.95 ±0.03
FinBERT (Yang et al., 2020)	85.63 ±0.02	86.64 ±0.02	73.91 ±0.04	67.35 ±0.03
BusinessBERT	**85.89** ±0.02	**87.02** ±0.02	**75.06** ±0.01	**67.71** ±0.01

over, both FinBERT models specifically trained for the financial domain perform worse than or the same as the general-purpose BERT-Base model.

Table 10: Results: Named entity recognition

	SEC filings		
	F1	Precision	Recall
BERT-Base	79.03 ±0.01	**79.80** ±0.01	78.87 ±0.03
FinBERT (Araci, 2019)	78.15 ±0.08	74.93 ±0.01	79.88 ±0.03
FinBERT (Yang et al., 2020)	78.25 ±0.02	75.64 ±0.04	81.93 ±0.02
BusinessBERT	**79.82** ±0.03	77.45 ±0.03	**83.38** ±0.01

4.3. Sentiment Analysis

BusinessBERT underperforms on the sentiment classification datasets (Financial Phrasebank, StockTweets) and in predicting continuous sentiment scores (FiQA). The results indicate a lack of polarity and opinions in BusinessBERT pretraining corpora. FinBERT (Araci, 2019) and BERT-Base display the most competitive performance among the sentiment analysis benchmark models. On average, BusinessBERT performs 5.79% worse than BERT-Base on sentiment analysis tasks.

4.4. Question Answering

In contrast to the other downstream tasks, the generation of reasoning programs and the calculation of financial ratios in question-answering are dissimilar; thus, this problem requires the encoder representations to adapt to new tasks. The results in **Table 12** summarize the performance of the question-answering benchmark models. BusinessBERT shows superior performance in predicting the correct answer (Exe Acc) and the correct mathematical derivation (Prog Acc)

Table 11: Results: Sentiment analysis

	FiQA		Financial Phrasebank		StockTweets	
	MSE	MAE	F1	Acc	F1	Acc
BERT-Base	0.0655 ±0.02	0.183 ±0.02	96.16 ±0.01	96.16 ±0.01	**72.33** ±0.04	**72.46** ±0.04
FinBERT (Araci, 2019)	0.0635 ±0.01	**0.176** ±0.01	**96.91** ±0.01	**96.91** ±0.01	71.53 ±0.04	71.62 ±0.04
FinBERT (Yang et al., 2020)	**0.0621** ±0.01	0.183 ±0.02	96.26 ±0.01	96.25 ±0.01	69.01 ±0.04	69.23 ±0.04
BusinessBERT	0.0758 ±0.02	0.202 ±0.02	96.08 ±0.01	96.07 ±0.01	69.14 ±0.06	69.54 ±0.05

in comparison with BERT-Base and the FinBERT models. On average, BusinessBERT performs 6.39% better than BERT-Base on question-answering tasks. Compared to BERT-Base, the pre-training approaches of both FinBERT models reduce the performance on the FinQA dataset.

Table 12: Results: Question answering

	FinQA	
	Exe Acc	Prog Acc
BERT-Base	56.06 ±0.01	54.14 ±0.01
FinBERT (Araci, 2019)	54.58 ±0.01	52.31 ±0.01
FinBERT (Yang et al., 2020)	46.64 ±0.01	44.90 ±0.01
BusinessBERT	**60.07** ±0.01	**57.19** ±0.01

5. Discussion of Findings

5.1. Corpora

We evaluate the performance of the business communication corpora individually by training a language model for each corpus. To investigate the contribution of the CompanyWeb corpus to the performance of BusinessBERT, we include a model variant trained only on the MD&A and S2ORC corpora. We follow the training procedure described in **Section 2** and display the evaluation results in **Table 13**.

Table 13: Ablation Studies: Corpora

	Risk		News		SEC filings			FiQA		Fin. Phrasebank		StockTweets		FinQA	
	F1	Acc	F1	Acc	F1	Precision	Recall	MSE	MAE	F1	Acc	F1	Acc	Exe Acc	Prog Acc
CompanyWeb	71.99	73.83	67.06	61.40	76.48	**86.81**	70.18	0.1356	0.291	84.32	84.67	48.22	49.62	47.25	46.03
MD&A	**84.46**	**85.51**	71.47	64.64	76.62	72.80	**81.35**	0.0830	0.216	**95.63**	**95.63**	67.86	67.92	54.49	52.83
S2ORC	82.64	83.93	72.45	65.14	**78.53**	77.73	80.25	**0.0823**	**0.214**	95.12	95.52	68.85	68.85	55.10	52.58
MD&A + S2ORC	82.04	83.38	**75.45**	**67.31**	68.01	70.80	69.09	0.0881	0.224	96.04	96.07	**69.80**	**69.85**	**56.15**	**54.58**

Table 14: Ablation Studies: IC

	Risk		News		SEC filings			FiQA		Fin. Phrasebank		StockTweets		FinQA	
	F1	Acc	F1	Acc	F1	Precision	Recall	MSE	MAE	F1	Acc	F1	Acc	Exe Acc	Prog Acc
BusinessBERT -IC	84.54	85.50	77.27	69.23	76.44	72.11	82.26	0.0771	0.201	96.16	96.16	69.36	69.46	57.80	56.41
BusinessBERT	85.89↑	87.02↑	75.06↓	67.71↓	79.82↑	77.45↑	83.38↑	0.0758↓	0.202↑	96.08↓	96.07↓	69.14↓	69.54↑	60.07↑	57.19↑

Table 15: Ablation Studies: Industry-Sensitive Fine-Tuning

	Risk		News		SEC filings			FiQA		Fin. Phrasebank		StockTweets		FinQA	
	F1	Acc	F1	Acc	F1	Precision	Recall	MSE	MAE	F1	Acc	F1	Acc	Exe Acc	Prog Acc
BERT-Base	82.89	84.58	71.41	65.12	79.03	79.80	78.87	0.0655	0.183	96.16	96.16	72.33	72.46	56.06	54.14
FT BERT-Base	83.39↑	84.85↑	71.62↑	65.46↑	78.00↓	78.86↓	77.88↓	0.070↑	0.191↑	95.19↓	95.14↓	73.92↑	73.62↑	57.45↑	55.54↑
RoBERTa-Base	84.97	85.35	74.97	68.42	79.59	80.31	79.29	0.0581	0.1675	97.81	97.79	79.42	78.72	56.10	48.00
FT RoBERTa-Base	85.74↑	87.30↑	75.33↑	63.43↓	79.32↓	78.71↓	80.46↑	0.0573↓	0.1644↓	97.76↓	97.74↓	77.82↓	77.65↓	49.96↓	48.65↑
LlaMA 2 7B	77.88	80.16	31.77	41.47	63.88	72.72	56.62	0.4101	0.4911	76.14	77.17	48.33	48.12	46.56	44.46
FT LlaMA 2 7B	79.22↑	81.64↑	41.08↑	47.14↑	65.37↑	73.36↑	58.55↑	0.3935↓	0.4646↓	78.83↑	79.47↑	49.72↑	49.41↑	48.56↑	46.73↑

We find that the models trained on the MD&A and S2ORC corpora, including the MD&A + S2ORC variant, outperform the model trained in the CompanyWeb corpus on all business NLP tasks. The performance gap is particularly evident for sentiment analysis benchmark models. Moreover, we do not observe remarkable performance differences between the MD&A and S2ORC models. Conversely, the model trained on both the MD&A and S2ORC corpora shows improved performance on the news, SEC filings, and FinQA benchmark datasets; however, this model performs similarly or worse than models trained on the individual MD&A and S2ORC corpora for other benchmark tasks.

5.2. Industry Classification

We evaluate the contribution of the IC pretraining objective to business NLP tasks by training a model variant using only the MLM and NSP objectives (BusinessBERT -IC). The results in **Table 14** show that the IC pretraining objective improves the downstream performance for the question-answering and named entity recognition tasks. The text classification results are unclear; BusinessBERT with the IC pretraining objective shows better performance on risk classification tasks, but underperforms on classifying news topics. For the sentiment analysis tasks, we observe mixed results for the FiQA and StockTweets datasets, while BusinessBERT -IC shows superior performance on the Financial Phrasebank dataset.

5.3. Industry-Sensitive Fine-Tuning

We explore the applicability of BusinessBERT's industry-sensitive pretraining approach to fine-tuning existing pretrained language models. Utilizing our business communication corpora and IC objective, we fine-tune BERT-Base (Devlin et al., 2019), RoBERTa-Base (Liu et al., 2019), and LLaMA 2 7B (Touvron et al., 2023) for a single epoch (Gururangan et al., 2020) [3]. Additional implementation details on the industry-sensitive fine-tuning are provided in **Appendix C**. The analysis results are presented in **Table 15**, where arrows indicate performance improvements and declines compared to the respective base model. For BERT-Base, the results in **Table 15** demonstrate that industry-sensitive fine-tuning yields performance benefits across similar business-related NLP tasks, although these benefits are less pronounced compared to full pretraining. Concerning

[3]The fine-tuned PyTorch checkpoints are available at `https://huggingface.co/pborchert/bert-ic`, `https://huggingface.co/pborchert/roberta-ic`, and `https://huggingface.co/pborchert/llama-ic-adapter`.

the RoBERTa model, industry-sensitive fine-tuning exhibits performance benefits solely on the Risk and FiQA datasets, with unclear results for the News and FinQA datasets. Conversely, for other benchmarks, model performance experiences a decline with the additional fine-tuning steps. In contrast, for the decoder-based LLaMA 2 7B model, industry-sensitive fine-tuning consistently enhances model performance across all business-related NLP tasks.

5.4. Vocabulary

To assess the efficiency of the BusinessBERT vocabulary, we encode business-related text using three vocabularies: BERT-Base (same as FinBERT (Araci, 2019)), FinBERT (Yang et al., 2020), and BusinessBERT. We conducted a comparative analysis of the token count required for text encoding in relation to BERT-Base, as illustrated in **Figure 6**. The results reveal that BusinessBERT and FinBERT (Yang et al., 2020) require fewer tokens to encode business-related text when compared to BERT-Base. Specifically, the vocabulary efficiency increased by 3.18% and 2.22% on average, respectively. These more efficient encodings of business-related text enable BusinessBERT to accommodate a greater amount of textual content within the maximum limit of 512 input tokens.

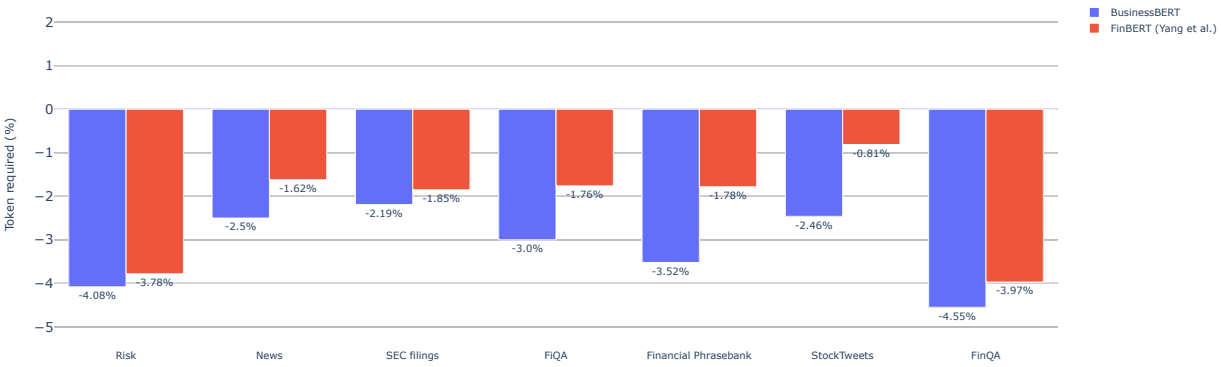

Figure 6: Vocabulary efficiency comparison. The Figure visualizes the ratio of token required to encode the benchmark datasets using the different vocabularies of the benchmark models. The ratios are computed with respect to the BERT-Base vocabulary, which serves as a baseline (0%).

6. Conclusion

In this paper, we improve the performance of language models for business-related NLP applications by introducing BusinessBERT, a new industry-sensitive language model. We start by calling attention to the added value of analyzing textual data for business and operations practice (see **Table 1**) and the challenges when textual data is analyzed through general-purpose language models. We design BusinessBERT to overcome these challenges by making two industry-sensitive adjustments to the BERT architecture (Devlin et al., 2019): (1) BusinessBERT is *trained on business communication corpora* consisting of company website content, MD&A statements and S2ORC papers in the business domain and (2) we include *industry classification as a pretraining objective* that enables BusinessBERT to capture industry-specific terminology. We contribute to OR practice in following ways:

(1) *Improved data-driven decision-making.* We demonstrate that BusinessBERT has superior performance leading to more informed decisions by decision makers. Our experiments show that BusinessBERT produces more accurate results than its benchmarks on text classification, named entity recognition and question-answering tasks. BusinessBERT shows average performance improvements between 1.24% and 6.39% over BERT-Base, between 2.14% and 9.69% over FinBERT (Araci, 2019) and between 0.71% and 28.08% over FinBERT (Yang et al., 2020). Furthermore, we demonstrate the transferability of our industry-sensitive language modeling approach to fine-tuning transformer encoders like BERT and RoBERTa, as well as decoder language models like LLaMA 2. Our results suggest that encoder models benefit from industry-sensitive fine-tuning on specific datasets, while industry-sensitive fine-tuning enhances LLaMA 2 performance across all business-related NLP tasks, resulting in an average performance improvement of 5.61%.

(2) *Reduced complexity.* Training language models is a time-consuming, computationally intensive and costly process for which organizations need to heavily invest in a specialized computing infrastructure, labeled training data and human expertise. This paper reduces this complexity by making BusinessBERT available to the OR community as a pretrained language model. Pretrained language models are off-the-shelf language models that are readily available to download and utilize. We make BusinessBERT freely available as a pretrained language model including its pretraining corpora and our benchmark datasets to replicate our results (`https://github.com/pnborchert/BusinessBERT`). BusinessBERT can be seamlessly implemented or

integrated into existing organizational (NLP) pipelines to achieve high performance quickly without too many fine-tuning efforts on new downstream tasks. This allows organizations to offload the high burden of heavy investments.

This paper inspires various paths for future research. Although we find that extending masked language modeling and next sentence prediction pretraining objectives with an industry classification objective improves downstream performance on business-related NLP tasks, we suggest that incorporating industry classification with other supervised, unsupervised or self-supervised objectives in multitask models as discussed by Raffel et al. (2020) is an interesting direction for future research. Despite great performance of BusinessBERT on text classification, named entity recognition and question-answering tasks, a limitation of BusinessBERT is displayed in its performance on sentiment analysis tasks. A possible explanation is the lack of polarized opinions in the pretraining corpora which requires further research.

By extending the BusinessBERT corpora with for instance analyst reports with corresponding industry segment labels, we can investigate the scalability of BusinessBERT's industry-sensitive pretraining approach. Additionally, future research opportunities extend to exploring industry-sensitive training approaches for a broader array of model architectures. Despite validating the transferability of our industry-sensitive training approach for fine-tuning transformer encoder and decoder models, it is crucial to acknowledge that these architectures are continually evolving, requiring ongoing development.

Acknowledgments

This research is supported by Google's TPU Research Cloud (TRC).

References

Aldunate, Á., Maldonado, S., Vairetti, C., Armelini, G., 2022. Understanding customer satisfaction via deep learning and natural language processing. Expert Systems with Applications 209, 118309. URL: https://www.sciencedirect.com/science/article/pii/S0957417422014397.

Alvarado, S., Cesar, J., Verspoor, K., Baldwin, T., 2015. Domain adaption of named entity recognition to support credit risk assessment, in: Proceedings of the Australasian Language Technology Association Workshop 2015, p. 84–90. URL: https://aclanthology.org/U15-1010.

Araci, D., 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063 [cs] URL: http://arxiv.org/abs/1908.10063.

Archak, N., Ghose, A., Ipeirotis, P.G., 2011. Deriving the pricing power of product features by mining consumer reviews. Management Science 57, 1485–1509.

Arts, S., Cassiman, B., Gomez, J.C., 2018. Text matching to measure patent similarity. Strategic Management Journal 39, 62–84. doi:10.1002/smj.2699.

Baechle, C., Huang, C.D., Agarwal, A., Behara, R.S., Goo, J., 2020. Latent topic ensemble learning for hospital readmission cost optimization. European Journal of Operational Research 281, 517–531.

Bao, Y., Datta, A., 2014. Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. Management Science 60, 1371–1391. URL: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2014.1930.

Bellstam, G., Bhagat, S., Cookson, J.A., 2021. A Text-Based Analysis of Corporate Innovation. Management Science 67, 4004–4031. URL: http://pubsonline.informs.org/doi/10.1287/mnsc.2020.3682, doi:10.1287/mnsc.2020.3682.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A Pretrained Language Model for Scientific Text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 3615–3620. URL: https://aclanthology.org/D19-1371, doi:10.18653/v1/D19-1371.

Beyer, A., Cohen, D.A., Lys, T.Z., Walther, B.R., 2010. The financial reporting environment: Review of the recent literature. Journal of Accounting and Economics 50, 296–343. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165410110000431.

Bhatia, S., 2019. Predicting Risk Perception: New Insights from Data Science. Management Science 65, 3800–3823. URL: https://pubsonline.informs.org/doi/10.1287/mnsc.2018.3121, doi:10.1287/mnsc.2018.3121.

Borchert, P., Coussement, K., De Caigny, A., De Weerdt, J., 2022. Extending business failure prediction models with textual website content using deep learning. European Journal of Operational Research URL: https://www.sciencedirect.com/science/article/pii/S0377221722005495, doi:10.1016/j.ejor.2022.06.060.

Bureau van Dijk, 2021. Orbis international company information. URL: https://orbis.bvdinfo.com/.

Chen, K., Li, X., Luo, P., Zhao, J.L., 2021a. News-induced dynamic networks for market signaling: Understanding the impact of news on firm equity value. Information Systems Research 32, 356–377. doi:10.1287/isre.2020.0969.

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.H., Routledge, B., Wang, W.Y., 2021b. FinQA: A Dataset of Numerical Reasoning over Financial Data, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. pp. 3697–3711. doi:10.18653/v1/2021.emnlp-main.300.

Choi, J., Menon, A., Tabakovic, H., 2021. Using machine learning to revisit the diversification–performance relationship. Strategic Management Journal 42, 1632–1661.

Davis, A.K., Piger, J.M., Sedor, L.M., 2012. Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language*. Contemporary Accounting Research 29, 845–868. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1911-3846.2011.01130.x, doi:10.1111/j.1911-3846.2011.01130.x.

Deng, S., Huang, Z.J., Sinha, A.P., Zhao, H., 2018. The Interaction Between Microblog Sentiment and Stock Returns: An Empirical Examination. MIS Quarterly 42, 895–918. doi:10.25300/MISQ/2018/14268.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for

Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) , 4171–4186.

Díaz, S.B., Coussement, K., Caigny, A.D., Pérez, L.F., Creemers, S., 2023. Do the us president's tweets better predict oil prices? an empirical examination using long short-term memory networks. International Journal of Production Research 0, 1–18.

Ewens, M., 2019. MD&A statements from public firms: 2002-2018 (Version: 1.0). URL: https://data.caltech.edu/records/1249, doi:10.22002/D1.1249.

Feldman, R., Govindaraj, S., Livnat, J., Segal, B., 2010. Management's tone change, post earnings announcement drift and accruals. Review of Accounting Studies 15, 915–953.

Feuerriegel, S., Gordon, J., 2019. News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. European Journal of Operational Research 272, 162–175.

Frankel, R., Jennings, J., Lee, J., 2021. Disclosure Sentiment: Machine Learning vs. Dictionary Methods. Management Science URL: https://pubsonline.informs.org/doi/10.1287/mnsc.2021.4156.

Geng, Z., Zhang, Y., Han, Y., 2021. Joint entity and relation extraction model based on rich semantics. Neurocomputing 429, 132–140.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 8342–8360. URL: https://aclanthology.org/2020.acl-main.740, doi:10.18653/v1/2020.acl-main.740.

Hartmann, J., Heitmann, M., Schamp, C., Netzer, O., 2021. The power of brand selfies. Journal of Marketing Research 58, 1159 – 1177. doi:10.1177/00222437211037258.

Hayes, P.J., 1992. Intelligent high-volume text processing using shallow, domain-specific techniques, in: Jacobs, P.S. (Ed.), Text-Based Intelligent Systems. Lawrence Erlbaum, Hillsdale, NJ.

Hoberg, G., Phillips, G., 2016. Text-Based Network Industries and Endogenous Product Differentiation. Journal of Political Economy 124, 1423–1465. URL: https://www.journals.uchicago.edu/doi/full/10.1086/688176.

Hong, J., Hoban, P.R., 2022. Writing more compelling creative appeals: A deep learning-based approach. Marketing Science 41, 941–965. doi:10.1287/mksc.2022.1351.

Hsu, M.F., Hsin, Y.S., Shiue, F.J., 2022. Business analytics for corporate risk management and performance improvement. Annals of Operations Research 315, 629–669.

Jeong, M., Minson, J., Yeomans, M., Gino, F., 2019. Communicating with warmth in distributive negotiations is surprisingly counterproductive. Management Science 65, 5813–5837. doi:10.1287/mnsc.2018.3199.

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2016. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 .

Katsafados, A.G., Leledakis, G.N., Pyrgiotakis, E.G., Androutsopoulos, I., Fergadiotis, M., 2024. Machine learning in bank merger prediction: A text-based approach. European Journal of Operational Research 312, 783–797. URL: https://www.sciencedirect.com/science/article/pii/S0377221723005982.

Koo, D.S., Julie Wu, J., Yeung, P.E., 2017. Earnings Attribution and Information Transfers. Contemporary Accounting Research 34, 1547–1579. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1911-3846.12308.

Lee, D., Hosanagar, K., Nair, H.S., 2018. Advertising content and consumer engagement on social media: Evidence from facebook. Management Science 64, 5105–5131.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics URL: https://doi.org/10.1093/bioinformatics/btz682, doi:10.1093/bioinformatics/btz682.

Li, F., 2010. The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach. Journal of Accounting Research 48, 1049–1102. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-679X.2010.00382.x, doi:10.1111/j.1475-679X.2010.00382.x.

Li, F., Wang, Z., Hui, S.C., Liao, L., Zhu, X., Huang, H., 2021. A segment enhanced span-based model for nested named entity recognition. Neurocomputing 465, 26–37. URL: https://www.sciencedirect.com/science/article/pii/S0925231221012911.

Li, Y., Xie, Y., 2020. Is a picture worth a thousand words? an empirical study of image content and social media engagement. Journal of Marketing Research 57, 1–19. doi:10.1177/0022243719881113.

Liu, A.X., Li, Y., Xu, S.X., 2021. Assessing the unacquainted: Inferred reviewer personality and review helpfulness. MIS Quarterly: Management Information Systems 45, 1113 – 1148. doi:10.25300/MISQ/2021/14375.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs] URL: http://arxiv.org/abs/1907.11692.

Liu, Y., Pant, G., Sheng, O.R.L., 2020. Predicting labor market competition: Leveraging interfirm network and employee skills. Information Systems Research 31, 1443–1466.

Lo, K., Wang, L.L., Neumann, M., Kinney, R., Weld, D., 2020. S2ORC: The semantic scholar open research corpus (version: 2020-07-05), in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 4969–4983. URL: https://www.aclweb.org/anthology/2020.acl-main.447, doi:10.18653/v1/2020.acl-main.447.

Loshchilov, I., Hutter, F., 2019. Decoupled Weight Decay Regularization, in: International Conference on Learning Representations (ICLR 2019). URL: http://arxiv.org/abs/1711.05101.

Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., Balahur, A., 2018. Www'18 open challenge: Financial opinion mining and question answering, p. 1941–1942. doi:10.1145/3184558.3192301.

Malo, P., Sinha, A., Takala, P., Korhonen, P., Wallenius, J., 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology , 782–796.

Miric, M., Jia, N., Huang, K.G., 2022. Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. Strategic Management Journal .

Moreno, A., Terwiesch, C., 2014. Doing business with strangers: Reputation in online service marketplaces. Information Systems Research 25, 865–886.

Mousavi, R., Raghu, T., Frey, K., 2020. Harnessing Artificial Intelligence to Improve the Quality of Answers in Online Question-answering Health Forums. Journal of Management Information Systems 37, 1073–1098. URL: https://www.tandfonline.com/doi/full/10.1080/07421222.2020.1831775, doi:10.1080/07421222.2020.1831775.

Narang, U., Yadav, M.S., Rindfleisch, A., 2022. The "idea advantage": How content sharing strategies impact engage-

ment in online learning platforms. Journal of Marketing Research 59, 61–78. doi:10.1177/00222437211017828.

Nauhaus, S., Luger, J., Raisch, S., 2021. Strategic Decision Making in the Digital Age: Expert Sentiment and Corporate Capital Allocation. Journal of Management Studies 58, 1933–1961. URL: https://onlinelibrary.wiley.com/doi/10.1111/joms.12742, doi:10.1111/joms.12742.

Netzer, O., Lemaire, A., Herzenstein, M., 2019. When words sweat: Identifying signals for loan default in the text of loan applications. Journal of Marketing Research 56, 960–980.

Pan, Y., Huang, P., Gopal, A., 2019. Storm clouds on the horizon? new entry threats and r&d investments in the u.s. it industry. Information Systems Research 30, 540–562. doi:10.1287/isre.2018.0816.

Puranam, D., Kadiyali, V., Narayan, V., 2021. The Impact of Increase in Minimum Wages on Consumer Perceptions of Service: A Transformer Model of Online Restaurant Reviews. Marketing Science 40, 985–1004. URL: https://pubsonline.informs.org/doi/10.1287/mksc.2021.1294, doi:10.1287/mksc.2021.1294.

Purda, L., Skillicorn, D., 2015. Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. Contemporary Accounting Research 32, 1193–1223. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1911-3846.12089, doi:10.1111/1911-3846.12089.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 21, 67.

Shi, Z., Lee, G.M., Whinston, A.B., 2016. Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence. MIS Quarterly 40, 1035–A53.

Shin, D., He, S., Lee, G.M., Whinston, A.B., Cetintas, S., Lee, K.C., 2020. Enhancing social media analysis with visual data analytics: A deep learning approach. MIS Quarterly 44, 1459–1492. doi:10.25300/MISQ/2020/14870.

Stevenson, M., Mues, C., Bravo, C., 2021. The value of text for small business default prediction: A deep learning approach. European Journal of Operational Research 295, 758–771. doi:10.1016/j.ejor.2021.03.008.

Sun, C., Wang, S., Zhang, C., 2021. Corporate Payout Policy and Credit Risk: Evidence from Credit Default Swap Markets. Management Science 67, 5755–5775. URL: http://pubsonline.informs.org/doi/10.1287/mnsc.2020.3753, doi:10.1287/mnsc.2020.3753.

Symitsi, E., Stamolampros, P., Daskalakis, G., Korfiatis, N., 2021. The informational value of employee online reviews. European Journal of Operational Research 288, 605–619. URL: https://www.sciencedirect.com/science/article/pii/S0377221720305269, doi:https://doi.org/10.1016/j.ejor.2020.06.001.

Taborda, B., de Almeida, A., Carlos Dias, J., Batista, F., Ribeiro, R., 2021. Stock market tweets data. URL: https://dx.doi.org/10.21227/g8vy-5w61, doi:10.21227/g8vy-5w61.

Theurer, C.P., Schäpers, P., Tumasjan, A., Welpe, I., Lievens, F., 2022. What you see is what you get? measuring companies' projected employer image attributes via companies' employment webpages. Human Resource Management 61, 543–561.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu,

Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Vairetti, C., Aránguiz, I., Maldonado, S., Karmy, J.P., Leal, A., 2024. Analytics-driven complaint prioritisation via deep learning and multicriteria decision-making. European Journal of Operational Research 312, 1108–1118. URL: https://www.sciencedirect.com/science/article/pii/S0377221723006562.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A.N., Gouws, S., Jones, L., Kaiser, L., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., Uszkoreit, J., 2018. Tensor2tensor for neural machine translation. CoRR abs/1803.07416. URL: http://arxiv.org/abs/1803.07416.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is All you Need, in: Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, G., Chen, G., Zhao, H., Zhang, F., Yang, S., Lu, T., 2021a. Leveraging multisource heterogeneous data for financial risk prediction: A novel hybrid-strategy-based self-adaptive method. MIS Quarterly 45, 1949–19998.

Wang, G., Chen, G., Zhao, H., Zhang, F., Yang, S., Lu, T., 2021b. Leveraging Multisource Heterogeneous Data for Financial Risk Prediction: A Novel Hybrid-Strategy-Based Self-Adaptive Method. MIS Quarterly 45, 1949–19998.

Wang, K., 2021. Is the Tone of Risk Disclosures in MD&As Relevant to Debt Markets? Evidence from the Pricing of Credit Default Swaps*. Contemporary Accounting Research 38, 1465–1501. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1911-3846.12644, doi:10.1111/1911-3846.12644.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A., 2020. Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online. pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6.

Xu, X., Qian, H., Ge, C., Lin, Z., 2020. Industry classification with online resume big data: A design science approach. Information & Management 57, 103182. URL: https://www.sciencedirect.com/science/article/pii/S0378720618307377.

Xu, Y., Armony, M., Ghose, A., 2021. The interplay between online reviews and physician demand: An empirical investigation. Management Science 67, 7344–7361.

Xu, Y., Tan, T.F., Netessine, S., 2022. The impact of workload on operational risk: Evidence from a commercial bank. Management Science 68, 2668–2693. doi:10.1287/mnsc.2021.4019.

Yang, Y., UY, M.C.S., Huang, A., 2020. FinBERT: A Pretrained Language Model for Financial Communications. arXiv:2006.08097 [cs] URL: http://arxiv.org/abs/2006.08097.

Zhang, M., Luo, L., 2022. Can consumer-posted photos serve as a leading indicator of restaurant survival? evidence from yelp. Management Science , mnsc.2022.4359.