

UNIVERSITY OF COPENHAGEN

DEPARTMENT OF ECONOMICS



SEMINAR: APPLICATIONS OF MACHINE LEARNING IN ECONOMICS

WORKFORCE SKILLS AS PERFORMANCE INDICATORS: ANALYZING FIRM JOB POSTINGS

JUNE 1, 2021

AUTHORS

VICTOR EMIL FUNCH

&

SEBASTIAN HONORÉ

Table of Contents

1	Introduction	2
2	Literature Review	4
2.1	Natural Language Processing and Vector Semantics	5
2.2	Human Capital and Firm Performance	8
3	Data	9
3.1	ESCO	11
4	Methodology	11
4.1	Categorizing Skills Using BERT	11
4.1.1	Pre-processing	12
4.1.2	The Transformer	12
4.1.2.1	Positional Encoding	12
4.1.2.2	Structure of the Encoder	14
4.1.3	Training and Implementation	15
4.2	Event Study	16
4.2.1	Identification of Events	17
5	Results	18
5.1	Classification	18
5.2	Event Study	20
5.3	Robustness	22
6	Conclusion and Limitations	22

1 Introduction

In dynamic and constantly changing markets, firms need the right set of capabilities to stay ahead of their competitors. One of the most strategic resources a firm can build on to achieve a sustainable competitive advantage is firm-specific knowledge ([Grant \(1991\)](#)). Consequently, sourcing workers with the right set of skills is a significant challenge for most businesses. The business needs to account for current skill needs and those that could potentially appear in the future if the workforce is not adequately prepared to meet future needs. A potential mismatch between firms skill requirements and the supply of skills may affect future firm performance through lower productivity or innovation (or both). Hence, the firm's market valuation will likely depend on its success in determining its future skill needs and attracting the right set of workers. As a result, investors should care about the investments firms make in specific workforce skills. This paper presents a novel method for measuring real-time demand for skill requirements and investigates whether or not the demand for skills is reflected in the current stock price utilizing an event study design.

Traditionally, three distinct types of skills have been examined concerning firm performance; Namely, science, technology, engineering and mathematics skills, in short STEM, creative skills and management skills. Research suggest that especially STEM skills is an enabling factor in firms' profitable innovation and future firm growth (see, e.g. [Leiponen \(2005\)](#), [Coad et al. \(2014\)](#)). Similarly, management skills have been linked with higher productivity, profitability, growth, survival rates, and innovation ([Bloom et al. \(2019\)](#)). Compared with the literature on STEM and management skills, there is relatively little evidence on the impact of creative skills on firm performance, and the evidence that exists is mixed. [Mollick \(2012\)](#) finds a positive effect of creative skills on firm performance. In contrast, [Siepel et al. \(2017\)](#) find that the use of creative skills alone has no impact on firm performance or innovation outcomes.

Despite these empirical findings, there is no consensus around the definitions of how to measure skills. Typically, there has been an equivalence between workers with a STEM background in a firm and the execution of STEM-related work tasks. When the relationship between workers with a STEM background and innovation is measured, it is unsure whether they are doing STEM work or utilizing STEM skills. Furthermore, classifying skills as creative poses an additional challenge as virtually all occupations require some creativity. Therefore, these skills have been proxied using occupational categories with a vital creative component, i.e. musicians, writers, producers, among others. Such proxies provide helpful information but are not sufficiently informed about the specific skills used in everyday work. In contrast, studies examining the effects of management skills have relied on qualitative approaches, which may fit the purpose better. It allows one to pinpoint specific skills employed within a firm. Nevertheless,

considering the speed at which financial markets move and the time it takes to retrieve qualitative data on skills, the information gathered will be obsolete before investors can use it. Furthermore, a general problem with the proxies used is that they rely on information concerning the current stock of skills in a firm. As a result, they only convey information on the current or past competitive advantage to which markets have already adjusted.

To examine the effects of specific workforce skills on firms performance in the financial markets, it is crucial to have access to real-time information on firms current skill demand and classify it accordingly. The current literature has only focused on the past stock of human capital employed by a firm contributing to the current firm performance. Hence, this methodology does not account for the present and future flow of human capital that can sustain or increase firm performance in the future. To the best of our knowledge, we are the first to examine the flow of human capital and its effect on firms financial performance.

We provide a novel method to pinpoint the exact skills demanded by firms instead of proxying them using educational backgrounds, occupations or gathering qualitative data. We utilize the rich source of data available through the online job portals JobIndex and JobNet, increasingly used by firms. Using Natural Language Processing (NLP), we can extract numerical representations of the information about specific skill requirements of a future employee conveyed in posted job-ads. Coupled with the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy, which provides definitions on 13.485 labour market skills, we can train a state-of-the-art NLP model to identify whether a given sentence in a job-ad contains a specific skill requirement and place it within the official ESCO taxonomy. We apply the tenet of NLP, namely, the Bi-directional Encoder Representations from Transformers (BERT), for the classification task at hand. A study by [Virtanen et al. \(2019\)](#) investigates different multilingual versions of BERT for a classification task. For a similar amount of training points, they report F1 scores between 74 to 82 percent. We expect that we can achieve similar results given the conceptual similarity of our task. Varying the shape of our training data to account for class imbalance, our best performing model achieves an F1 score and accuracy of 71 and 79 percent, respectively (see table 2).

We find that job-ads demanding workers with STEM skills positively and significantly affect a firm's stock price. Job-ads demanding workers proficient working with computers have a 0.0011 percent effect, whereas information skills have an effect of 0.0058 percent on the stock price. Both the magnitude and the sign of the effects are as expected. The magnitude is arguably small, but since we estimate the effect of one employee on expected future firm performance, it seems reasonable that the effect is

minimal. We estimate that those job-ads demanding workers with creative and vocational skills do not affect expected future firm performance. However, management skills have a positive and significant effect of 0.00017 percent on the stock price. The effect of management skills is small compared to the literature examine the effect of CEO changes on the stock price (see, e.g. [Bonnier and Bruner \(1989\)](#)). However, the job-ads we consider demanding management skills are likely to be middle managers, project managers, and internal recruiters, which have a lesser effect on future firm performance.

Although online job-ads offer a rich and detailed view of skill requirements, some limitations may be present. The sample of job-ads may not represent all job openings because of differences in recruitment practices by occupation and industry. Jobs requiring high-level skills may be more likely to be advertised online, whereas low-skilled work is not. On the other hand, jobs requiring very high-level skills such as CEO's and other c-suite skills may not be listed as these are typically found through job scouts. Moreover, the skills included in a job-ad description may not be a complete listing of skills required for the job. Some skills may be tacit within the job-ad for some occupational categories. Those included may reflect what differentiates the job at hand from similar positions or reflect the employee's general values. Another problem likely to exist is some degree of measurement errors related to job-ads. Some ads may leave the number of jobs available unspecified, which implies that we do not measure the effect of a single employee. Furthermore, measurement errors may arise due to discrepancies between what categories are annotated by the annotators. Lastly, we need to take into account that the model is never more accurate than 79 percent with an F1 score of 71 percent.

The rest of the paper is structured as follows: in section 2 we provide a review of the literature governing Natural Language Processing and the theoretical link between human capital and firm performance. In section 4 we present the BERT model and present the empirical method used to examine the link between workforce skills and financial performance. The data used in this paper is described in section 3. The empirical results are presented in section 5. Finally, section 6 concludes.

2 Literature Review

This section introduces concepts of NLP and the economic theory connecting human capital with firm performance. The first part of the literature review is concerned with the theory of vector semantics and how the numerical representation of words can incorporate linguistic properties. The essence of NLP is to quantify text data within a corpus (collection of texts which in our case is the database of job-ads). The second part of the literature review presents the traditional human capital theory and the resource-based view of the firm, where knowledge and skills play a crucial role in achieving competitive

advantage and hence, increase in stock price.

2.1 Natural Language Processing and Vector Semantics

NLP is a branch of computer science and artificial intelligence where natural language is understood as human communication. Thus, the starting point of NLP is the study of language (linguistics). Within linguistics, semantics examines the relationship between words and their meaning. It seeks to uncover how words are interrelated and can create meaning such that clear communication arises, whether by conversation or text. When words within the same semantic field are communicated to a recipient, they create a clear underlying understanding or meaning of these words. A semantic field is a set of words that cover a particular semantic domain and bear structured relations with each other (Jurafsky and Martin (2020)). For example, the co-appearance of the words "python", "java", "programming", "CPU", "script" in a text will point to the semantic field of *computer science*.

In NLP, vector semantics are used to represent the meaning of words, analyze semantic fields, and quantify and structure text data. The goal of vector semantics is to position a target word in a multidimensional semantic space given its neighboring words, also called the *context words*. The idea of representing and analyzing words in a multidimensional space goes back to Switzer (1965) and was referred to as mechanical indexing. Following the conventions of NLP, we will refer to vector representations of words as *embeddings*.

In the seminal paper by Mikolov et al. (2013), the authors present a revolutionary method for computing continuous vector representations of words (also referred to as Word2Vec). Their novelty is twofold; they obtain short, dense embeddings instead of long and sparse embeddings; they learn embeddings using unlabeled data. At the time of publication the embeddings obtained by Mikolov et al. (2013) "... work better in every NLP task than sparse vectors" (Jurafsky and Martin (2020)). Mikolov et al. (2013) suggests two algorithms for learning embeddings. We will touch upon the algorithm with the highest semantic score in the paper: the skip-gram algorithm with negative sampling. This way to learn embeddings was revolutionary as the skip-gram algorithm is used on running text. This implies that the training process is a self-supervised task. This is advantageous when working with text data as it is generally easy to access but always unlabeled. Firstly, the skip-gram algorithm reads text sequentially and treats the target word (the one to predict) and its neighboring words as a positive sample. It then randomly select other neighboring words from the corpus and treats these as a negative sample. Secondly, a logistic classifier is trained to predict the probability of the target word appearing in the context of its neighboring words. Lastly, the learned weights of the classifier are saved and used as embeddings. This

results in two embeddings for one word. One for the word as a target and one for the word as a context word. These embeddings are notably shorter than what was previously the case and compress much more information about the word and its relatedness to other words. Mikolov et al. (2013) showcase the semantic properties of the embeddings by simple linear algebra. Let x denote a word and let $v(x)$ denote the embedding for that word. The authors show that the vector z ,

$$z = v(king) - v(man) + v(woman), \quad (1)$$

is closest to the embedding for "queen", $v(queen)$, using the cosine similarity measure defined as

$$\text{cosine}(a, b) = \frac{a \cdot b}{|a||b|} \quad (2)$$

which is the dot product normalized by the length of the vectors. Hence, the more information is within the same dimensions of the embedding, the higher the dot product. Conversely, embeddings that have a dot product of zero will be orthogonal and thereby unrelated. Despite the obvious power of Word2Vec embeddings, they have the drawback of being static embeddings. This means that embeddings returned by the algorithm are fixed for the given training corpus. This implies that the corpus biases the embeddings and that the embeddings obtained by Mikolov et al. (2013) are somewhat non-contextual for applications outside of the corpus. For example, homonyms will have the same embedding despite their difference in meaning.

In NLP, there are several ways to overcome the problem of static embeddings. Cho et al. (2014) introduce what is now called the sequence-to-sequence model. The model consists of an encoder and a decoder, two separate Recurrent Neural Networks (RNN). The encoder RNN takes a sequence (sentence) as input say $x = (x_1, x_2, \dots, x_t)$ and reads each element (word) sequentially while updating the hidden layer h_t according to

$$h_t = f(h_{t-1}, x_t) \quad (3)$$

where f can be any non-linear activation function, e.g. the Sigmoid. After reading the input sequence x , the hidden layer summarises the elements within x . The input sequence is thus transformed into a new and lower-dimensional representation of x say z , which is a fixed-length vector (Cho et al. (2014)). If the input sequence is a vector of word embeddings representing a text sequence, then passing it through the encoder RNN means that z will represent both the context and the embeddings. The decoder then takes the encoder output z and turns it back into a language representation.

Using the sequential nature of the text and the dependency between words solved the problem of static embeddings being non-contextual however, this is also the Achilles heel of the model. It does not allow encoding sequences in parallel, meaning that only one embedding is computed at a time. This leads to

computational inefficiency when dealing with longer sequences and a large amount of data. Furthermore, the approach proposed in [Cho et al. \(2014\)](#) is forcing the RNN to squeeze all information of a sequence into a fixed-length vector. As a result of this property, the follow-up paper by [Cho et al. \(2014\)](#) shows that it is difficult for the model to handle long sequences. Especially sequences that are longer than the sequences in the training corpus.

Firstly, [Bahdanau et al. \(2014\)](#) address the problem of the RNN encoder-decoder that is forcing information into a fixed-length vector. They suggest extending the model with an attention mechanism. The most crucial distinguishing feature of this approach from the previously described is that it does not attempt to encode a whole input sequence into a single fixed-length vector. Instead, it encodes the input sentence into a sequence of vectors and chooses a subset of these vectors adaptively. The authors show that this extension outperforms its predecessor. Secondly, [Vaswani et al. \(2017\)](#) address the computational inefficiency of training sequential models when computing contextual embeddings. In their seminal paper, they introduce the Transformer architecture based solely on the attention mechanisms and does not involve an RNN. It, too, consists of an encoder and a decoder part. The authors find the Transformer to be *"be superior in quality while being more parallelizable and requiring significantly less time to train"* ([Vaswani et al. \(2017\)](#)). Instead of using the sequence to create context, the Transformer use positional encoding, self-attention and a simple Feed Forward Neural Network (FFNN). The positional encoding keeps track of each element in the sequence while the attention mechanisms forward a complete picture of the whole sequence through layers of the model.

In the light of the Transformer [Devlin et al. \(2019a\)](#) introduces the BERT model. As the name suggests, Bidirectional Encoder Representations from Transformer consists of the encoder part of the Transformer architecture in [Vaswani et al. \(2017\)](#). This implies that BERT can encapsulate the context of a sequence without comprising computational speed. In section 4.1 the encoder of the Transformer and thereby BERT is explained in depth. [Rogers et al. \(2020\)](#) review the current uses of BERT by surveying 150 papers in which the model is applied and finds that BERT is the go-to model for NLP applications. This is further confirmed by the General Language Understanding Evaluation (GLUE) benchmark leader board. GLUE is a collection of resources for training, evaluating, and analyzing natural language understanding models. As of April 24, 2021, four models in the top five are BERT versions (see appendix A1). Therefore, BERT seems the obvious choice for the task at hand: reading and classifying skill requirements in job-ads.

2.2 Human Capital and Firm Performance

The theoretical literature on the link between human capital and firm performance goes back to [Becker \(1964\)](#). The *classical* theory posits that investments in human capital can increase firm productivity through knowledge and skill gains. Nevertheless, these productivity gains are not expected to improve firms financial performance, at least in a perfectly competitive labour market. An individual's wage rate is determined by her marginal productivity in other firms in such a market. A firm can only capture human capital returns if it increases the marginal product of labour more than the wage rate. Though, today it is generally recognized that labour market frictions exist, implying that labour markets exhibit some degree of imperfect competition (see, e.g. [Alan \(2011\)](#)). The notion of rent and whom it accrues is at the core of labour market imperfections. Most notably is the case where rents accrue to the firm. In this case, the marginal product is above the wage rate, and investments in human capital positively affect firm performance. In sum, the existence of imperfect labour markets generates opportunities for firms to reap the financial benefits of human capital investments.

Similarly, the relationship between financial performance and human capital has been studied within the managerial resource-based theory (RBT) (see, e.g. [Crook et al. \(2011\)](#)). According to RBT, firms who possess valuable resources that their competitors cannot easily imitate will achieve what is known as a *competitive advantage* ([Barney \(1991\)](#)). Inherently this leads to the conclusion that observed differences in firm performances can be accounted for through differences in resource distributions. Traditionally, firm resources include all assets, capabilities, organizational processes, firm attributes, information and knowledge controlled by a firm ([Daft and Lengel \(1983\)](#)). These resources enable firm's to conceive and implement strategies that increase their performance. The entire spectrum of resources capable of optimizing performance is enormous. Therefore the RBT literature has focused on three separate entities: Physical capital resources ([Williamson \(1975\)](#)), human capital resources ([Becker \(1964\)](#)) and organizational capital resources ([Tomer \(1987\)](#)). Of the three entities human capital has received the most widespread attention([Coff \(1997\)](#), [Grant \(1991\)](#), [Crook et al. \(2011\)](#)). Human capital may be a valuable resource capable of achieving competitive advantage because it may be in short supply and semi-permanently tied to the firm. As such, competitors cannot purchase the resources and compete away any advantage a firm may have ([Peteraf \(1993\)](#)). Hence, from a resource-based perspective, human capital is widely regarded as one of the key drivers of financial performance arising from the competitive advantages created.

While the theory predicts a positive relationship between human capital and firm performance, the empirical literature has provided mixed results. The concepts of human capital and firm performance are

broad and related to the empirical question at hand. Departing from overall effects In a meta-analysis [Newbert \(2007\)](#) considers seven articles on the relationship. The seven articles include 33 individual hypotheses tests of the human capital relationship as discussed above, 11 (33 percent) of which were supported. Similarly, [Crook et al. \(2011\)](#) conducted a meta-analysis of 66 articles with different human capital and firm performance constructs. The authors concluded that there likely exists a positive empirical relationship as expected from the theoretical models.

Most of the literature on the relationship have dealt with performance in terms of accounting measures, i.e. return on Assets (ROA), return on equity (ROE), sales growth and profitability, among others (see, e.g. [Lopez \(2003\)](#), [Tanriverdi and Venkatraman \(2005\)](#), [Lee et al. \(2001\)](#)). When measuring firm performance in accounting terms, only the past or current competitive advantage is captured. In contrast, few have dealt with the forward-looking *sustainable* competitive advantage grounded in finance measures such as stock returns ([Barney \(2014\)](#)). [Riley et al. \(2017\)](#) examines the effect of human capital investments on the market valuation of firms using an event study methodology. The sample consisted of 99 publicly traded business units with a total of 219 events. An event occurred if a business unit received an award for human capital investments. The authors found that a signal of firms' effective investments in human capital leads to a cumulative abnormal return of 1.67 percent. Although the study examines the effects on a firm's forward-looking sustainable competitive advantage, the events only account for the current stock of human capital contributing to current firm performance. Hence, it does not account for the flow of present and future human capital that can sustain this performance in the future. To the best of our knowledge, we are the first to examine the flow of human capital and its effect on financial performance.

3 Data

In this paper, we analyze danish firms that have been publicly traded on Nasdaq Copenhagen in the years between 2007 and 2020. The sample consists of 97 firms covering all sectors of the Danish economy. The financial data was gathered from the Yahoo Finance API. Given that job-ads are posted daily, we opted for daily stock data. The returns we analyze are all based on the adjusted closing price for each stock.

The job-ads used in this paper are mainly retrieved from JobIndex, and a small fraction originates from JobNet. In total, we have a database consisting of roughly 3.7 million job-ads spanning the period between January 1, 2007, and April 31, 2020.

For the classification task, we need labelled training data. Based on sentences in job-ads marked as skill requirements, we label the sentences following the official ESCO taxonomy (explained in section 3.1). The annotation of categories is done at the second skill level (see appendix A2), and we have obtained the feature sentence and the target label. This amounts to 10,715 training points. Furthermore, each of the 13,485 skills in ESCO is accompanied by sentences describing the skills. These sentences bear the same structure as those in the job-ads, and we add them to our training data. However, we avoid using them in the test set as we want to predict only on job-ads (further elaborated in section 5.1). Combining annotated and ESCO data, we end up with 24,560 training points in our data. We note that the training data suffers from a class imbalance as 20 percent of the categories have less than 100 training points.

After training our model, we use it to predict skill requirements and enrich the 3.7 million job-ads with an ESCO category. We have obtained data that enables us to investigate if the demand for skills is reflected in the stock price. Specifically, we consider the highest possible aggregation of skills for the event study design, including eight unique skill categories (see appendix A2). Table 1 presents these skill categories along with the number of events, defined as the posting of a job-ad highlighting this specific skill. The event is determined by the most frequent skill in the job-ad. In total, our dataset includes 62,765 events with the skill category "Communication, collaboration and creativity", making up 51 percent of the event count.

Table 1: Number of events within skill categories

Skills	Events
Construction	62
Assisting and caring	112
Communication, collaboration and creativity	32,392
Working with computers	12,851
Handling and moving	354
Information skills	12,856
Management skills	3,678
Working with machinery and specialized equipment	460
Total	62,765

3.1 ESCO

ESCO is a multilingual classification or taxonomy of skills, competencies, qualifications and occupations relevant for the European labor market, education and training. The skills pillar of ESCO contains 13485 skills classified in a hierarchical structure.

The current literature examining the effects of skills on firm performance have focused on STEM, creative and management skills. The ESCO taxonomy allows for a more disaggregated and detailed approach than currently employed in the literature. We can classify STEM skills into two categories: "Working with computers" and the second being "Information Skills". The category "Communication, collaboration and creativity" on the upper ESCO level is arguably broad. However, the subcategories strongly emphasize creativity, and we perceive this category as mainly creative skills. Management skills are neither disaggregated further at the top level, and we classify management as its own type. Finally, in addition to the three classical skills typically analyzed, ESCO also allows us to estimate vocational skills' effects. As these types of skills are often practical, we perceive "Construction", "Assisting and caring", "Handling and moving", and "Working with machinery and specialized equipment" as being part of the overall category vocational skills.

However, the ESCO classification is not always clear-cut. For example, "Working with computers" includes subcategories that usually appear in some definitions of the creative industries, which might entail that these should be classified as creative skill. Nonetheless, we take the stance that working with computers is generally considered a core STEM skill and classified as such (see, e.g. the O*NET definition). This highlights one of the challenges when annotating sentences. This ambiguity might lead annotators to categorize the same skill into different categories.

4 Methodology

First, this section introduces the BERT model in-depth. Namely, how input sequences are pre-processed and how contextual embeddings are produced within the model's layers. Secondly, we propose an event study design to measure the effect of specific workforce skills on firm market valuation.

4.1 Categorizing Skills Using BERT

We use the BERT model, which builds upon the Transformer architecture, to compute embeddings and classify skill requirements within the official ESCO taxonomy. First, we consider the pre-processing of input sequences needed in any NLP application before passing to the model. Secondly, we review the encoder part of the Transformer architecture as this essentially what makes up BERT. Thus, we will go

through positional encoding and a single encoder unit structure as BERT is a stack of encoders. Lastly, this section will comment on implementing BERT as it comes pre-trained by Google Research and needs to be fine-tuned for downstream tasks.

4.1.1 Pre-processing

The first step in pre-processing is to break the unstructured sequence of words into tokens. BERT uses the WordPiece tokenizer, an algorithm trained to learn merge rules of all characters in the training corpus such that a vocabulary is formed. For example, "working" will become "work"+"ing" as the probability of the merged character pairs of "work" and "ing" respectively is higher than "working". For a more detailed explanation, we refer to [Wu et al. \(2016\)](#).

The sequence is then turned into a numerical representation by replacing each token with its index value from the vocabulary. Each index value represents the tokens one-hot vocabulary vector and is scaled to be the same length. If the sequence before scaling is shorter than the chosen length, it is padded. The length of a sequence is a hyperparameter set by the researcher, and we set the maximum sequence length equal to 100 as no sentence in our corpus is longer.

4.1.2 The Transformer

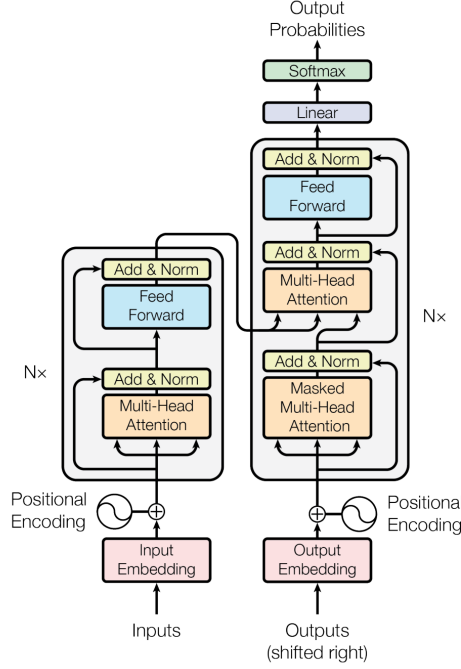
As touched upon in section 2.1 the Transformer is based on the sequence-to-sequence model but breaks with the sequential nature of an RNN-encoder. The Transformer does not read a sequence left to right or right to left. Instead, it reads the whole sequence at once and is said to be non-directional or *bi-directional*. The architecture of the Transformer is shown in Figure 2 as initially presented in [Vaswani et al. \(2017\)](#).

The left-most gray box in Figure 2 is the encoder part. The "Nx" symbolizes that it is a stack of encoders. In the original paper, it is a stack of $N_x = 6$ encoders, whereas BERT is a stack of $N_x = 12$ encoders. Before we look at the encoder structure, we will first consider the positional encoding that keeps track of the elements in a given sequence and makes parallelization possible. The following sections are structured such that they follow the flow of the encoder (The left-most gray box in Figure 2). Above we described the tokenization, and our inputs are ready to be handled. Hence, the following sections will describe positional encoding, then turn to the multi-head attention, and comment on the Feed Forward Neural Network (FFNN).

4.1.2.1 Positional Encoding

After pre-processing, positional encoding is applied to keep track of the position of the embeddings in a sequence. For a given embedding, a positional vector with the same dimension is computed and then

Figure 1: Architecture of the Transformer



Source: [Vaswani et al. \(2017\)](#)

added to the embedding such that a new and enriched embedding is returned. A simple positional encoding PE could be mapping $x_t \rightarrow t$ in a sequence $x = (x_1, x_2, \dots, x_t)$ such that the $PE = (1, 2, \dots, t)$. However, the scale of the numbers will be large for long sequences, which can pose many problems, e.g., exploding gradients when passed on to the neural network. An immediate solution to this problem would be making PE relative by dividing the sequence length. However, this poses another problem. A value of 0.2 in a PE vector provides different information in a sequence length of 5 compared to a sequence with a length of 20.

To obtain embeddings enriched with information about the relative and absolute position of the embedding in the given sequence, which overcomes the issues described above, we follow [Vaswani et al. \(2017\)](#) and compute the positional encoding as

$$\begin{aligned} PE_{(t, 2i)} &= \sin\left(\frac{t}{10000^{2i/d}}\right) \text{ and,} \\ PE_{(t, 2i+1)} &= \cos\left(\frac{t}{10000^{2i/d}}\right) \end{aligned} \quad (4)$$

where t is the position of the embedding, i is the element i th element within the embedding vector and d is the dimension of the embedding. The authors show that computing each the position of the embedding using a sinusoidal function and varying the wavelength based on the dimension produces the same results as using learned positional embeddings, but that the sinusoidal works better for sequences longer than

sequences encountered in the training corpus.

4.1.2.2 Structure of the Encoder

BERT consists of $Nx = 12$ identical encoders, each of which has two layers. The multi-head self-attention mechanism and a feed-forward neural network. Firstly, we will review self-attention and multi-headed attention. Lastly, we will comment briefly on the role of the neural network.

The goal of the self-attention mechanism is to weigh the sequence of embeddings to capture the similarity between them. Recall that the cosine in equation (2) captures the similarity between vectors. Hence, computing the dot-product between a given embedding x and all the other embeddings in the sequence will return a weight for each embedding for x . Multiplying the weights with the other embeddings will return a new embedding with the same dimensionality as the original embedding x . Nevertheless, x has been re-weighted towards itself, given its similarity with the other embeddings. The idea of self-attention can be illustrated by the idea of smoothing time series data. This can be done by re-calculating a given observation by a weighted combination of neighboring observation and itself. This can be seen as computing self-attention based on proximity, whereas we for the text-data base it on similarity. Note that this process has no weights or hyper-parameters that can be learned during a training process. It is, however, easily introduced into the concept of self-attention.

Following Vaswani et al. (2017) and Devlin et al. (2019a) we use the scaled dot-product attention as our attention mechanism. It is defined as

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (5)$$

where Q, K and V denote query, key and value respectively and are weight matrices will that will be learned, d is the embedding dimension and softmax is the normalized exponential function defined as

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^K x_j} \text{ for } i = 1, 2, \dots, K \text{ and } x = x_1, x_2, \dots, x_K. \quad (6)$$

The attention mechanism works the following way. Take a given embedding of dimension $1 \times n$ and multiply it with the query matrix Q of dimension $n \times n$. In the first encoder layer, this matrix is created by random initialization. This yields a query vector of dimension $1 \times n$. Now compute the dot-products (similarity) with all the other embeddings or keys, i.e. we multiply the query vector with the key matrix K ($n \times n$). We have now obtained a weight matrix ($1 \times n$) for re-weighting the embeddings, which we for numerical stability normalize by the square root of the embedding dimension d (Vaswani et al. (2017)). Lastly, take the softmax of the normalized weights and multiply with the value matrix V ($n \times n$). The value matrix is consisting of original embeddings received by the encoder. We have now obtained a

vector of dimension $1 \times n$. This is the attention vector with the same shape as the original embedding passed to the encoder, but it has now been re-weighted towards itself, given the other embeddings in the sequence. The softmax transformation ensures that all weights are within $[0,1]$.

How are we sure that one attention vector is capturing enough information? Key and value matrices are computed in parallel in the multi-head attention layer to address this multiple query. Multi-head attention is defined as

$$\text{multi-head}(Q, K, V) = \text{concat} (a(Q_1, K_1, V_1)_1, a(Q_2, K_2, V_2)_2, \dots, a(Q_h, K_h, V_h)_h,) \quad (7)$$

where a refers to the attention mechanism in equation (5). In the original Transformer, there are $h = 8$ heads of attention where $h = 12$ in BERT. Note how each self-attention head has its own weight matrices to compute; thus, the layers do not share weights. For each layer, an attention vector is computed, then concatenated into a final vector. Since we concatenate attention vectors from 12 layers we now have a vector of dimensionality $1 \times (12n)$. Therefore, we pass it to the second and final layer of the encoder.

The attention vector returned from the multi-headed attention layer is passed to a fully connected FFNN. This is essentially just a multi-layer perceptron, and the architecture will not be covered in this paper (for a detailed explanation, see [Goodfellow et al. \(2016\)](#)). The purpose of the FFNN is twofold. Firstly, the FFNN is designed such that the weight matrix of the FFNN will project the concatenated attention vector into the original dimension space of the input. Thus, the weight matrix is of dimension $(12n) \times (12n)$ such that the output and final embedding is $1 \times n$. This implies that the attention vector is turned into an acceptable form by the next encoder in the stack or at the output layer. Secondly, the FFNN will learn patterns within and between sequences of attention vectors as the weights are kept and updated throughout the training process. Thus, combining attention and pattern recognition will result in contextual embeddings.

4.1.3 Training and Implementation

The training process of BERT is done in two steps called pre-training and fine-tuning, respectively. The goal of pre-training is to generate a general language understanding model by training on a large corpus. It is done by applying Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). The advantage of pre-training is that few parameters need to be learned from scratch, and BERT is loaded with trained weights (e.g. from Google Research’s GitHub repository ([Devlin et al. \(2019b\)](#))). Thereby, we avoid tuning around 110 million hyper-parameters. Fine-tuning aims to take the general model and adjust it to the vocabulary within job-ads and apply it to a downstream task (classification).

The pre-training process is an unsupervised learning task, and therefore BERT can be trained on a vast corpus. The Danish BERT is trained on Google Books and the Danish Wikipedia. The steps of training (MLM and NSP) are applied simultaneously. MLM is applied to learn context within a sequence and mask 15 percent of the tokens in the sequence. The objective is then to predict the original token of the masked word based only on its surrounding tokens and, thereby, context. NSP is applied to learn the relationship or context between sequences in the corpus. The NSP method will take two adjacent sentences, X and Y , and then draw a random sentence Z from the corpus. Z will replace Y 50 percent of the time, and the binary task of predicting what comes after X is trained. Thus, the unsupervised training data consists of masked tokens and pairs of originally adjacent or randomly paired sentences.

Fine-tuning is the process where pre-trained weights and hyper-parameters from [Devlin et al. \(2019b\)](#) are loaded. Hence, we use a stack of $N_x = 12$ encoders and $h = 12$ multi-head attention layers. Then an untrained neural network with only one layer is added at the end of the encoder stack, and the fine-tuning process can begin. Our labelled training data is passed through the model where contextualized embeddings are computed and classified according to the ESCO taxonomy. In the training process, the AdamW learning rate optimization algorithm is applied to find the right hyperparameters of the neural network. AdamW is based on stochastic gradient descent and is described in [Kingma and Ba \(2014\)](#). This will update weights in the neural network and the rest of the model architecture such that BERT "learns the language" of skill sentences.

Thus, we have arrived at a method suited for our goal of categorizing skill requirements in unstructured text from online job-ads. We can tokenize sentences and create meaningful embeddings. Together with our labelled training data, we can train a neural network to classify sentences of skill requirements according to the official ESCO classification.

4.2 Event Study

To quantify whether the firm's market valuation is affected by the demand for specific workforce skills, we use an event study design. A crucial assumption underlying the research design is that capital markets are sufficiently efficient to react to events, i.e. job-ads demanding specific skills ([Malkiel and Fama \(1970\)](#)). That is, current stock prices adjust to the release of all new public information. We take the stance that job-ads is publicly available information as they are accessible online. The stock return behavior ultimately depends on the investor's expectation regarding the specific skills and how this adds to future firm performance. The logic is that the objective of a firm is to maximize profits which depends on the revenues generated and costs incurred, including the opportunity cost of capital. Financial investors

measure the forward-looking sustainable competitive advantage and firm performance by net present value when setting the economic market value of a firm. The value of the stock issued by the firm should therefore vary with expectations regarding future firm performance. The firm's total value should equal the discounted value of the future cash flows, and the value of a firm's equity should equal the discounted value of the future cash flows of the entire firm (less expected payments to the owners of the firm debt). The stock price of the firm should reflect the discounted value of these future cash flows. If a specific skill is associated with increased marginal productivity and the firm accrues rents, this leads to a positive net present value of the job-ad. Thus, the firm generates a future cash flow which should be reflected in the stock price today.

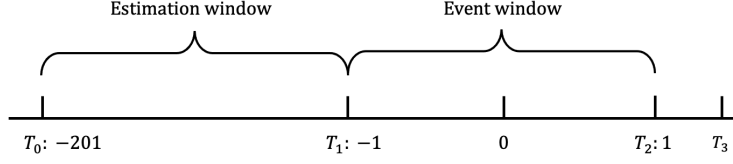
4.2.1 Identification of Events

The event is the posting of an online job-ad. The type of event is categorized based on the primary skill required from the employer. In total, an event can belong in one of eight skill categories based on the ESCO classification as described in section 3. The specific event date is the same day as the job-ad was posted. We do not have the exact timing of job-ad postings, allowing for a more exact intra-day effect. Observations for firms who posted more than one job-ad on a given day has been removed from the sample. The event window we consider is very narrow and spans one day before and one day after the event has occurred. We use this event window because some firms post job-ads daily, and as such, disentangling the effects would be very difficult over a longer event horizon. Therefore, we restrict our results to be presented non-graphically. Having only three points of time, one day before the event, the event and the day after, do not produce a meaningful figure. A solution could be extending the event window; however, this contradicts the logic of choosing the narrow window. In section 5.3 we check the robustness of event window specification.

Before we can estimate any abnormal stock returns, we need to measure the normal returns over some period of time. The usual estimation window of normal returns spans 100-300 days prior to the event (see e.g., [Armitage \(1995\)](#)). When choosing the estimation window, we face a trade-off between greater precision in our estimation of normal returns and these returns becoming more "out of date". As such, we decided to use a 200 days estimation window. However, the results are robust to this specification; for a graphical presentation, see figure 2.

Several methods to derive the normal rate of returns has been proposed in the literature. The simplest model is to assume that a given stock will earn the market rate of return $R_{i,M_{i,t}}$ within the estimation window considered. Abnormal returns are then the actual return less the market return. The most common approach is to estimate a stock's normal returns through a regression on the market returns

Figure 2: Estimation and event window



and is known as the *market model*. Formally, we estimate:

$$R_{i,t} = \alpha_i + \beta_i R_{i,M_{i,t}} + \eta_{i,t} \quad (8)$$

Where $R_{i,t}$ is the return of stock i on trading day t , α_i and β_i are the regression coefficients of the daily return rate of the stock i and the market return rate. Knowing the normal returns we can find the abnormal returns in the event window by:

$$AR_{i,t} = R_{i,t} - (\alpha_i + \beta_i R_{i,M_{i,t}}) \quad (9)$$

The cumulative abnormal return (CAR) of stock i in the event window is then:

$$CAR_{i,(t_1,t_2)} = \sum_{t_2}^{t=t_1} AR_{i,t}. \quad (10)$$

CAR allows us to capture the cumulative effect of the posting of a job-ad. This may be spread over several days surrounding the event day because of the gradual availability of information and interpretation of the event's impact on future firm profitability.

5 Results

First, we present our results of the classification of skills and secondly, we present the results of applying the event study and test the robustness of the event window.

5.1 Classification

The results from the training process are presented in table 2. The table shows the top five results in the descending order of the F1 score from our model selection process. The best performing model and the one we will use to identify events have an F1 score of 72 percent. It also turns out to be the specification that yields the highest accuracy. We choose our model based on the F1 score. When evaluating classification tasks, one cannot blindly use the accuracy, which is simply a measure of how many predictions were right out of all predictions. Thus, the measure is not adequate in the presence of false positives and false negatives. Therefore we calculate the precision and recall of the model. Precision measures how many positive instances are true positives, i.e. how many instances are predicted correctly and recall measures how many of the true positives were correctly identified. The F1 score is then the harmonic

Table 2: Classification results

Random State	Scenario	N	Epochs	Accuracy	F1 Score	Precision	Recall	Test Size
3	3	170	6	0.79	0.71	0.72	0.71	2097
3	3	140	6	0.78	0.7	0.71	0.69	2097
1	4	-	6	0.76	0.64	0.64	0.64	2456
2	4	-	6	0.74	0.59	0.58	0.61	2456
1	2	170	6	0.79	0.55	0.56	0.54	2097

Note: The training has 24560 observations and is consisting of annotated job-ad data and official ESCO data.

mean of precision and recall and will be a better metric in our case with class imbalances.

Besides using AdamW to find the optimal parameters, we do something else to select our model. The prediction task is only meant to be applied to job-ads, so we define four scenarios for our training data, as it is a combination of annotated data and ESCO sentences. This further implies that we do not include ESCO data in our test set, i.e. we only use it as complementary training data. In a sense, we introduce our own hyperparameters to tune, which determines the shape of the training data. The first scenario is using only the annotated data. The second scenario is annotated data and ESCO data in small categories (this refers to N in table 2 and is categorized with less than N training points). The third is annotated data in large categories (equal or larger than N) and ESCO data in small categories. Finally, the fourth scenario is combining both data sources such that all data is used. In scenario 2 and 3, we allow the category size cut-off N to vary across 100, 140, 170 and 210. These are chosen somewhat arbitrarily but kept in mind that 20 percent of the categories have less than 100 training points. We also choose to tune the epoch of the classifier to mitigate over- and under-fitting of the model. Epochs are the number of times that the entire training data is passed through the model in batches. We vary the epoch from one to ten and do not go above as each increase in epoch becomes increasingly computationally expensive and might lead to overfitting. Lastly, we vary the random state of the test-train split to ensure no split in which categories with few observations are not represented in the test set or that we consequently split in a bad-performing manner. We only let the random state take four different values to limit the time of the training.

For the five best performing specification, the epoch amounts to six. This suggests that going below six epochs will not allow the model to learn the patterns in the training data sufficiently, whereas going above will over-fit the model. More interestingly, we find that the best performing scenario is scenario 3, with 170 and 140 as the cut-off value, respectively. This indicates that when there are only a

few training points, the particular sentences within the ESCO corpus contain essential information that is else diluted. Using the entire training data yields similar accuracy, but a lower F1 score and results for larger N is not in the top 5.

Thus, we select our training data based on scenario 3 with N equal to 170. We set the epoch to 6 and obtained the highest F1 score of 71 percent and the highest accuracy of 79 percent. We apply this model to the database of job-ads.

5.2 Event Study

As discussed above, our main research question asks whether the firm market valuation is affected by the demand for specific workforce skills. Table 3 presents CAR for the eight ESCO skill categories. The event window considered is from $[-1,1]$ as described in section 4.2. The results indicate that job-ads demanding STEM skills positively and significantly affect the firm's market valuation. Job-ads looking for workers proficient working with computers have a 0.0011 percent effect on the stock price, whereas information skills have an effect of 0.0058 percent. It seems reasonable that information skills have a larger effect as this type of skill is generally much more complex and more in line with traditional STEM skills. It includes skills such as "conducting studies", "analyzing and examining data", and "performing mathematical calculations", to name a few. These results imply that investors' expectations regarding the demand for STEM skills are that they positively affect future firm performance. However, the effects are small, which is somewhat expected since the marginal effects of one employee on firm future performance are likely small.

Although no studies have estimated the effect of STEM skills on firms' market valuation, some have found positive effects of this type of skills on firm performance. Siepel et al. (2019) finds that firms who actively have invested in employing workers with STEM skills have a 10.1 percent higher turnover growth than firms who did not. Similarly, Leiponen (2005) finds that evidence of a positive relationship between the share of STEM graduates in a firm and firm performance mediated by a higher level of innovation.

The effect of creative skills is very close to zero and insignificant, implying that investors have no expectations of higher future firm performance from such workers. A similar result is found in Siepel et al. (2016). The authors examine the effects of creative skills on various measures of firm performance, including employment growth, sales growth, productivity and innovation growth. In all of their specifications, they find no significant effect on the use of creative skills and firm performance. In contrast, Brunow et al. (2018) find a significant effect on firm-level innovation from the employment of creative

Table 3: Event study results

Skills	CAR %
STEM skills	
Working with computers	0.0010986*** (0.0003702)
Information skills	0.00579*** (0.0002895)
Creative skills	
Communication, collaboration and creativity	0.0000002610 (0.0002322)
Vocational skills	
Constructing	-0.0799303 (0.0597924)
Assisting and caring	0.0020559 (0.0040058)
Handling and moving	-0.0040297* (0.00226)
Working with machinery and specialized equipment	-0.0031273 (0.0023318)
Management skills	
	0.0001667*** (0.00006858)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust std. errors in parenthesis

skills.

All of the vocational skill categories but Handling and moving are insignificant. Handling and moving have a -0.004 percent effect on the firm's market value, significant at 10 percent. The effect of vocational skills on firm performance has not been studied previously in the literature. These types of skills are very related to the type of business under consideration, whereas STEM and creative skills are much more general. Consider the case of the Danish shipping company, A. P. Møller Mærsk. While the company rely on STEM, creative and managerial skills, a large part of their core business relies on the employment of workers with vocational skills. They need workers capable of operating their ships, handling and transporting the containers. Therefore, one would expect that increasing the demand for these types of

skills in business' where their core business relies on them would be perceived as a signal of improved future firm performance. However, as it turns out, vocational skills do not affect the stock price on average.

Management skills have a positive and significant effect of 0.00017 on the stock price. The positive link between management skills and firm performance has been studied extensively and in various practices. However, typically only the effect of CEO or top management changes has been studied concerning the stock market. [Bonnier and Bruner \(1989\)](#) disentangle the "news" effect from the real effect, i.e. the expectation regarding future performance. They estimate that changes in top management have a real effect on the stock price of 2.5 percent.

In comparison, our estimate is relatively small. The reason might be that the sample we consider does not include job-ads for either top management or CEO's as these are typically found through job scouts. As a result, the job-ads we consider demanding management skills is likely to be middle managers, project managers, internal recruiters, among others.

5.3 Robustness

In this section, we check the robustness of our estimates by varying the length of the event window. Table 4 presents the CARs for event windows of $[-5,5]$, $[-2,2]$ and the baseline $[-1,1]$. As noted in section 4.2 we choose the $[-1,1]$ event window as this will likely be least contaminated by other events. The further we move from this, we should expect to contaminate the results with the effects of other job-ad postings.

Overall the conclusion that STEM skills and management skills have a positive impact resides. Similarly, the conclusion of no effect of creative and vocational skills is robust across specifications. The magnitude of the effects is generally also larger at wider event windows. This is likely a contamination effect.

At event windows of $[-5,5]$ and $[-2,2]$, the STEM skill Working with computers becomes insignificant and Working with machinery and specialized equipment becomes significant. Constructing is only significant at the $[-2,2]$ specification.

6 Conclusion and Limitations

In this paper, we introduce a novel method to measure the real-time demand for skill requirements of Danish firms utilizing online job-ads and state-of-the-art NLP. This enables us to estimate the effect of specific workforce skills on firms financial performance measured by the stock price. Contrary to previous studies, our method overcomes various limitations caused by restricted data collection techniques. We can classify skills in over 3.7 million job-ads and link these to changes in the stock price. We achieve

Table 4: Alternate event windows

Skills	[-5,5]	[-2,2]	[-1,1]
STEM skills			
Working with computers	-0.000591 (0.0006709)	-0.00058283 (0.00045600)	0.0010986*** (0.0003702)
Information skills	0.0016075*** (0.0005973)	0.00149065*** (0.00042394)	0.00579*** (0.0002895)
Creative skills			
Communication, collaboration and creativity	0.0001133 (0.0004272)	0.00036498 (0.00029040)	0.0000002610 (0.0002322)
Vocational skills			
Constructing	-0.0186874 (0.0172605)	-0.02072062* (0.01203992)	-0.0799303 (0.0597924)
Assisting and caring	0.002575 (0.0068829)	0.00075832 (0.00547367)	0.0020559 (0.0040058)
Handling and moving	0.0017823 (0.0049002)	-0.00258698 (0.00294792)	-0.0040297* (0.00226)
Working with machinery and specialized equipment	-0.0083266* (0.0043601)	-0.00788014*** (0.00293322)	-0.0031273 (0.0023318)
Management skills			
	0.0024527*** (0.0012555)	0.00023956*** (0.00006816)	0.0001667*** (0.00006858)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust std. errors in parenthesis

reasonably high accuracy of the classification task of 79 percent and an F1 score of 71 percent. Using an event study methodology, we find that job-ads demanding STEM skills have a positive and significant effect on the firm's market valuation. Job-ads demanding workers proficient working with computers have a 0.0011 percent effect, whereas information skills have an effect of 0.0058 percent on the stock price. We estimate those job-ads demanding workers with creative and vocational skills do not affect expected future firm performance. However, management skills have a positive and significant effect of 0.00017 percent on the stock price.

There are, however, some caveats in our application of NLP. Firstly, the predictions of our model are never going to be more accurate than 79 percent, implying that every fifth skill is wrongly classified

(and possibly even more considering the F1 score). Secondly, the annotation of training data is done by hand, which might induce human errors. Some categories will have similar features such that doubt arises about which category is the right one, e.g. "developing games" can be both a computer skill and a creative skill. The outcome of the annotators might differ in situations like these. However, this is less of a concern as the annotation is done on the second level, and the categories are clearer. Finally, we suspect that job-ads are biased towards the category "Communication, collaboration and creativity" as this is an omnipresent skill. Recall that this constitutes 51 pct of the events. "Working with others", "Solving problems", "Obtaining information verbally" (see Appendix A2) is demanded no matter the profession. The bias then arises if much weight is put on these skills within the job-ads opposed to the skills specific to the actual job. The challenge is that the specific skills are tacit and nested in the job title. For example, the job title "nurse" nests many specific and fundamental skills already obtained. The employer might be looking for an individual that possesses supplementary skills within other ESCO categories. Thus, our identification of events is subject to measurement errors, suggestive of further development of the training data for future research purposes.

There is another problem of potentially *contaminating* events within the event window, i.e. news distinct from the job-ads, which may affect the share price. However, [Thompson \(1988\)](#) demonstrate that unrelated news should have zero effect on the share price on average in a large sample. Furthermore, we posit that the posting of job-ads is not systematically linked with other types of events such as acquisitions, share issues, loans, to name a few.

References

- Alan, M. (2011). Imperfect competition in the labor market. In *Handbook of labor economics*, Volume 4, pp. 973–1041. Elsevier.
- Armitage, S. (1995). Event study methods and evidence on their performance. *Journal of economic surveys* 9(1), 25–52.
- Bahdanau, D., K. Cho, and Y. Bengio (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of management* 17(1), 99–120.
- Barney, J. B. (2014). *Gaining and sustaining competitive advantage*. Pearson higher ed.
- Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago press.
- Bloom, N., E. Brynjolfsson, L. Foster, R. Jarmin, M. Patnaik, I. Saporta-Eksten, and J. Van Reenen (2019). What drives differences in management practices? *American Economic Review* 109(5), 1648–83.
- Bonnier, K.-A. and R. F. Bruner (1989). An analysis of stock price reaction to management change in distressed firms. *Journal of Accounting and Economics* 11(1), 95–106.
- Brunow, S., A. Birkeneder, and A. Rodríguez-Pose (2018). Creative and science oriented employees and firm-level innovation. *Cities* 78, 27–38.
- Cho, K., B. Van Merriënboer, D. Bahdanau, and Y. Bengio (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.
- Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Coad, A., M. Cowling, P. Nightingale, G. Pellegrino, M. Savona, and J. Siepel (2014). Innovative firms and growth: Uk innovation survey.
- Coff, R. W. (1997). Human assets and management dilemmas: Coping with hazards on the road to resource-based theory. *Academy of management review* 22(2), 374–402.

- Crook, T. R., S. Y. Todd, J. G. Combs, D. J. Woehr, and D. J. Ketchen Jr (2011). Does human capital matter? a meta-analysis of the relationship between human capital and firm performance. *Journal of applied psychology* 96(3), 443.
- Daft, R. L. and R. H. Lengel (1983). Information richness. a new approach to managerial behavior and organization design. Technical report, Texas A and M Univ College Station Coll of Business Administration.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019a). BERT: Pre-training of deep bidirectional transformers for language understanding.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019b). Google research’s BERT repository: <https://github.com/google-research/bert>.
- Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio (2016). *Deep learning*, Volume 1. MIT press Cambridge.
- Grant, R. M. (1991). The resource-based theory of competitive advantage: implications for strategy formulation. *California management review* 33(3), 114–135.
- Jurafsky, D. and J. H. Martin (2020). *Speech & Language Processing*. Draft of December 30, 2020.
- Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Lee, C., K. Lee, and J. M. Pennings (2001). Internal capabilities, external networks, and performance: a study on technology-based ventures. *Strategic management journal* 22(6-7), 615–640.
- Leiponen, A. (2005). Skills and innovation. *International Journal of Industrial Organization* 23(5-6), 303–323.
- Lopez, V. A. (2003). Intangible resources as drivers of performance: Evidences from a spanish study of manufacturing firms. *Irish Journal of Management* 24(2), 125.
- Malkiel, B. G. and E. F. Fama (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance* 25(2), 383–417.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space.
- Mollick, E. (2012). People and process, suits and innovators: The role of individuals in firm performance. *Strategic Management Journal* 33(9), 1001–1015.






- Newbert, S. L. (2007). Empirical research on the resource-based view of the firm: an assessment and suggestions for future research. *Strategic management journal* 28(2), 121–146.
- Peteraf, M. A. (1993). The cornerstones of competitive advantage: a resource-based view. *Strategic management journal* 14(3), 179–191.
- Riley, S. M., S. C. Michael, and J. T. Mahoney (2017). Human capital matters: Market valuation of firm investments in training and the role of complementary assets. *Strategic Management Journal* 38(9), 1895–1914.
- Rogers, A., O. Kovaleva, and A. Rumshisky (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics* 8, 842–866.
- Siepel, J., R. Camerani, and M. Masucci (2019). Skills combinations and firm performance. *Small Business Economics*, 1–23.
- Siepel, J., R. Camerani, M. Masucci, and G. Pellegrino (2016, May). The fusion effect: the economic returns to combining arts and science skills. Project report.
- Siepel, J., M. Cowling, and A. Coad (2017). Non-founder human capital and the long-run growth and survival of high-tech ventures. *Technovation* 59, 34–43.
- Switzer, P. (1965). Vector images in document retrieval. *Statistical association methods for mechanized documentation*, 163–171.
- Tanriverdi, H. and N. Venkatraman (2005). Knowledge relatedness and the performance of multibusiness firms. *Strategic management journal* 26(2), 97–119.
- Thompson, J. E. (1988). More methods that make little difference in event studies. *Journal of Business Finance & Accounting* 15(1), 77–86.
- Tomer, J. F. (1987). *Organizational capital: The path to higher productivity and well-being*. Praeger Pub Text.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need.
- Virtanen, A., J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo (2019). Multilingual is not enough: Bert for finnish.
- Williamson, O. E. (1975). Markets and hierarchies: analysis and antitrust implications: a study in the economics of internal organization. *University of Illinois at Urbana-Champaign’s Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Appendices

A1 GLUE Leaderboard

GLUE Leaderboard April 24, 2021

Rank	Name	Model	URL	Score
1	ERNIE Team - Baidu	ERNIE		90.9
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8
3	HFL iFLYTEK	MacALBERT + DKM		90.7
	Alibaba DAMO NLP	StructBERT + TAPT		90.6
	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6

Source: <https://gluebenchmark.com/leaderboard>.

Accessed: April 24, 2021.

A2 ESCO Overview

ESCO skill taxonomy at upper and second level

Upper skill level	Second skill level
Construction	Building and repairing structures Installing interior or exterior infrastructure Finishing interior or exterior of structures
Assisting and caring	Counselling Providing health care or medical treatments Protecting and enforcing Providing information and support to the public and clients Preparing and serving food and drinks Providing general personal care
Communication, collaboration and creativity	Negotiating Liaising and networking Teaching and training Presenting information

(To be continued)

Upper skill level	Second skill level
	Advising and consulting Promoting, selling and purchasing Obtaining information verbally Working with others Solving problems Designing systems and products Creating artistic, visual or instructive materials Writing and composing Performing and entertaining Using more than one language
Working with computers	Programming computer systems Setting up and protecting computer systems Accessing and analyzing digital data Using digital tools for collaboration, content creation and problem solving Using digital tools to control machinery
Handling and moving	Sorting and packaging goods and materials Moving and lifting Transforming and blending materials Tending plants and crops Assembling and fabricating products Making moulds, casts, models and patterns Using hand tools Positioning materials, tools or equipment Handling animals Cleaning Washing and maintaining textiles and clothing Handling and disposing of waste and hazardous materials
Information skills	Conducting studies, investigations and examinations

(To be continued)

Upper skill level	Second skill level
	Documenting and recording information
	Managing information
	Processing information
	Measuring physical properties
	Calculating and estimating
	Analysing and evaluating information and data
	Monitoring, inspecting and testing
	Monitoring developments in area of expertise
Management skills	Developing objectives and strategies
	Organising, planning and scheduling work and activities
	Allocating and controlling resources
	Performing administrative activities
	Leading and motivating
	Building and developing teams
	Recruiting and hiring
	Supervising people
	Making decisions
Working with machinery and specialized equipment	Operating mobile plant
	Driving vehicles
	Operating watercraft
	Operating machinery for the extraction and processing of raw materials
	Operating machinery for the manufacture of products
	Using precision instrumentation and equipment
	Installing, maintaining and repairing mechanical equipment
	Installing, maintaining and repairing electrical, electronic and precision equipment
	Operating aircraft