Bridging the Gap: Implementing Large Language Models for Precision-Centric Startup Sector Classification in Sparse Data Environments

Stanislav Kharchenko
twotensor.com
daniel@twotensor.com

Abstract

1 Introduction

The dynamic landscape of venture capital hinges significantly on the effective alignment of potential investment projects with a fund's defined thesis profile. Among a multitude of factors that constitute this alignment, discerning the industry sector of prospective investment opportunities is particularly challenging, yet unequivocally crucial.

Traditionally, this task has been entrusted to skilled human analysts, a process that involves deep-diving into each investment opportunity to extract pertinent information (Kaplan and Stromberg, 2001). While human analysis has been lauded for its accuracy and in-depth understanding, it is also associated with considerable time investment, financial expenditure, and inherent limitations when it comes to scalability, which has led to calls for a more efficient, automated methodology (Hochberg, 2016).

The advent of Large Language Models (LLMs), like OpenAI's GPT-3 (Brown et al., 2020) or Google's T5 (Raffel et al., 2019), presents a promising avenue for such automation. These advanced models have the ability to analyze textual data autonomously, a capability that is particularly pertinent when evaluating descriptions of prospective investment projects. Their primary function is to interpret these descriptions and accordingly categorize the associated projects into predefined industry sectors without human supervision.

Despite their promising capabilities, LLMs have not yet convincingly demonstrated superior performance over human analysis in this classification task. A key stumbling block is the prevalence of ambiguities and omissions in many project descriptions, which hinders the accurate extraction of industry-specific information. In contrast, human analysts can assimilate a broader array of data, like visual cues and other contextually relevant information.

In this work, we traverse a spectrum of state-of-the-art models and machine-learning techniques that have exhibited proficiency in the field of text classification and generation. The core proposition of our study is a synergistic ensemble approach that melds these diverse strategies, achieving an exemplary feat of surpassing human performance when operating within the constraints of sparse data environments prevalent in the venture capital domain.

2 Background

Large Language Models (LLMs) have fundamentally reshaped the field of Natural Language Processing (NLP). This transformation is evidenced in pioneering models such as Google's Text-to-Text Transfer Transformer (T5) (Raffel et al., 2019), OpenAI's Generative Pretrained Transformer (GPT) series (Radford et al., 2018; Radford et al., 2019; Brown et al., 2020), and the innovative model, XLNet (Yang et al., 2019). These sophisticated models transform language into machine-understandable representations, effecting a revolution in tasks such as translation, question-answering, and text classification.

OpenAI's GPT series employs a transformer-based architecture (Vaswani et al., 2017) and uses unsupervised learning to generate contextually relevant text. The latest iteration, GPT-4, surpasses a trillion parameters and can produce text that is nearly indistinguishable from human-written prose. The GPT series has found extensive applications across various domains, including the categorization of venture capital projects.

Conversely, Google's T5 adopts a unified text-to-text approach, reframing all NLP tasks as text-generation problems. This strategy enables T5 to perform a multitude of tasks without the need for task-specific model architectures, thereby enhancing its versatility in text processing.

Further, XLNet (Yang et al., 2019), a model developed by researchers at Google Brain and Carnegie Mellon University, predicts the probability of a word given all other words in a sequence. This attribute enhances its understanding of context and could potentially improve the precision of venture capital project categorization.

The advent of LLMs, such as T5, GPT, and XLNet, suggests the possibility of automating tasks traditionally performed by humans, thanks to their ability to understand and generate human-like text. This has profound implications for various industries, including venture capital, where the potential of LLMs to analyze and categorize project descriptions could significantly streamline operations and improve the efficiency of industry matching.

Building on this, the recent work by Cao et al., (2022), outlines a scalable and adaptive system that leverages the prompt and model tuning of generative LLMs to infer the industry sectors of companies. Their approach combines the strengths of advanced LLMs with detailed fine-tuning strategies to create an effective system that outperforms traditional rule-based and machine-learning methods in classifying companies into industry sectors. While the precision of their results subceeds our estimates of human classification capabilities, their

methodology offers valuable insights for our research direction.

Despite these advancements, challenges persist in the implementation of LLMs. Ambiguities and incomplete information often found in project descriptions continue to pose significant obstacles for automated analysis (Taddy, 2015; Bao and Datta, 2014). Traditionally, human analysts have leveraged visual data and other contextual information to bridge these gaps, but achieving this level of complex inference with LLMs remains a challenge. Overcoming these issues and maximizing the potential of LLMs to match or even surpass human performance in venture capital sector categorization is the primary focus of our research.

3 Method

The core technique in our study is centered around adopting a two-stage, task-specific fine-tuning strategy, harnessing a generative model, specifically T5, and a discriminative model, namely XLNet, to create a streamlined pipeline for industry classification. The methodology, rooted in the recent findings of Cao et al., (2022), indicates a minimal performance variation between large and extra-large model sizes with respect to total precision and recall, thereby leading us to opt for the large-sized T5 model.

In the initial phase, the T5 model is fine-tuned to take a textual description of a startup, supplemented by contextual details such as the credentials of the founding team, market size, competitive landscape, and funding history, and expand upon this information by generating additional context-specific text. The fine-tuning operation involves training the model on an extensive corpus of analogous startup descriptions, ultimately empowering it to generate text that significantly enhances the input details.

Proceeding to the next phase, the XLNet model, lauded for its aptitude in context understanding and text sequence interpretation, is fine-tuned for a multi-task classification scheme across predefined industry sectors. The model is trained on a large-scale dataset, composed of startup descriptions that have been expertly classified into industry sectors. This rigorous training facilitates the model's recognition of subtle linguistic patterns and sector-specific terminologies, thus enabling it to classify startups into their corresponding sectors accurately.

The challenges, including handling of ambiguous and incomplete information, as identified in earlier works (Taddy, 2015; Bao and Datta, 2014), form a crucial aspect of our training regimen. The objective is to equip the model to surmount these hurdles and deliver accurate predictions, thereby closely emulating or even surpassing the classification performance of human analysts.

In essence, our methodology unfolds as a bifurcated process: the generation of detailed startup descriptions using a fine-tuned T5 model, followed by a precise industry sector classification of these descriptions via a fine-tuned XLNet model. This approach seeks to synergize the capabilities of both generative and discriminative models, aspiring to enrich the description of startups and

heighten the precision and efficiency of the sector classification task.

Algorithm 1 Fine-tuning

Require: Startup descriptions and contextual details in the form of $d \rightarrow c$, a generative NLP model $P(c|d;\theta_1)$, discriminative NLP model $P(s|c;\theta_2)$, generative model freezing steps t', learning rates ϵ_1 and ϵ_2

Ensure: The optimal parameters θ_1^* and θ_2^*

Initialize θ_1 by loading the pretrained T5 model
Initialize θ_2 by loading the pretrained XLNet model
for $t = 1$ to T **do**
    Sample a mini-batch from the startup descriptions and contextual details
    Transform each d into an expanded description c' by feeding it through the fine-tuned T5 model
    Forward propagate c' to obtain the sector prediction $\hat{s}$ using the XLNet model
    Calculate the cross-entropy loss $L(\hat{s}, s)$
    if $t \leq t'$ **then**
        $\epsilon = \epsilon_1$
    else
        $\epsilon = \epsilon_2$
        $\theta_1 := \theta_1 - \epsilon \nabla_{\theta_1} L(\hat{s}, s)$
    end if
    $\theta_2 := \theta_2 - \epsilon \nabla_{\theta_2} L(\hat{s}, s)$
    $\theta_1^* = \theta_1$ and $\theta_2^* = \theta_2$
end forreturn θ_1^* and θ_2^*