Textual Representation of Business Plans and Firm Success

Maria Saveria Mavillonio*

Abstract

In this paper, we leverage recent advancements in deep learning models to extract information from business plans on various equity crowdfunding platforms and predict the success of firm campaigns. Our approach spans a broad and comprehensive spectrum of model complexities, ranging from standard textual analysis to more intricate textual representations - e.g. Transformers-, thereby offering a clear view of the challenges in understanding of the underlying data. To this end, we build a novel dataset comprising more than 640 equity crowdfunding campaigns from major Italian platforms. Through rigorous analysis, our results indicate a compelling correlation between the use of intricate textual representations and the enhanced predictive capacity for identifying successful campaigns.

Keywords: Crowdfunding, Text Representation, Natural Language Processing, Transformers

JEL Classification: C45, C53, G2, G23, L26

*The research acknowledges funding support from the PRIN grant no. 20177FX2A7, provided by the Italian Ministry of University and Research.
Department of Economics and Management, University of Pisa, Italy.
Email: mariasaveria.mavillonio@phd.unipi.it

1 Introduction

Textual analysis of firm information disclosure has demonstrated an increasing ability to forecast firm behavior and performance (Loughran and McDonald, 2011; Loughran and McDonald, 2016). For instance, Cohen, Malloy, and Nguyen (2020) proved how investors might overlook subtle yet crucial signals in annual reports that significantly influence stock prices.

Among all the ready available firms' documents that could be analysed, we focus on the Business Plan (BP). This is the primary source of information when startups, companies, spin-offs, one-person businesses and even researchers in academia, would like to showcase a new idea and seek approval and funding to finance it. Hence, BP serve the purpose of generating, disseminating, and exchanging knowledge with significant stakeholders, including investors, financial institutions, governmental bodies, prospective collaborators, and incubators or accelerators (Hormozi et al., 2002; Delmar and Shane, 2003). Consequently, the conversion of knowledge into the business plan and subsequent dissemination of these concepts to external stakeholders emerge as pivotal concerns and strategic operational processes for nascent enterprises. In practice, BP contains a detailed description of the proposed business, market and industry analysis, production plans, operations and logistic details, customer analysis, financial analysis, competitive environment, and other details(McKenzie and Sansone, 2019).

One notable application of BP is crowdfunding, especially equity-based crowdfunding. Crowdfunding is an innovative method of collecting capital for a new company activity from the general public instead of using traditional methods like bond issuance or bank lending. It can be hard to determine the intrinsic value of a project considering the recent increase in the volume and amount of funding asked through crowdfunding campaigns. Indeed, crowdfunding disrupts traditional financing like bond issuance or bank lending. Thus, identifying successful startups in advance is crucial for investors seeking optimal returns and directing capital to the most promising projects. It's also essential for governments targeting programs at high-growth firms and researchers studying successful entrepreneur characteristics. Additionally, if factors predicting high growth can be influenced, this could drive policy efforts to enhance these attributes in individuals lacking them.

In this context, the literature can be divided into two macro strands; one investigates key factors in the success of a crowdfunding campaign, and the other predicts post-campaign outcomes (see Deng et al. (2022) for a review). In general, most of the approaches pursue the goal by focusing on the campaign characteristics (e.g. the number of backers, duration, overfunding, goal, category, etc.). Less attention has been paid to other sources of

information such as text, visuals, social networks, updates, and comments associated with the campaign.

A notable exception is Zhou et al. (2018), who study the process by which project owners raise funds from backers, allowing to identify the determinants of the campaign's success from the text description (e.g. length, tone and sentiment). On the same line, Kaminski and Hopp (2020) apply machine learning techniques to predict the outcome of crowdfunding startup pitches using text, speech, and video metadata, emphasising the need to understand crowdfunding from an investor's perspective.

Signori and Vismara (2018) study a small population of successfully funded equity project funding that the degree of investor participation predicts post-campaign success. Finally, McKenzie and Sansone (2019) study companies' success in Nigerian business competition by analysing their business plans. Interestingly they find that human evaluators are no better than an ML algorithm in predicting this success. However, most of this research does not accurately assess the quality of the business idea being unable to capture all the elements of the business project, including the BP.

Algorithmic text analysis is gaining prevalence due to the increasing availability of large-scale corpora, a trend expected to persist with growing text data accessibility. However, economists lack consensus on how to optimally employ text algorithms due to their novelty, resulting in the absence of a unified methodology (Ash and Hansen, 2023; Athey and Imbens, 2019). Moreover, the rapid development of Deep Learning models, particularly Transformers, has transformed text analysis, surpassing prior assessments of text-as-data methods in economics (Gentzkow, Kelly, and Taddy, 2019) and in social science (Ziems et al., 2024). Transformers excel in detecting subtle patterns and semantic meanings in language, creating a succinct vectorial representation of the input text (Vaswani et al., 2017)[1].

Although the obtained representation from transformers lacks interpretability, we believe that it is worth investigating their ability to extract and encode impact variables from economic documents (e.g. business plans). On the one hand, this would help to have a clearer view of the capacity of the deep learning models in this context. On the other hand, it would provide a strong baseline that should be considered in the future to assess interpretable models.

To achieve this, we compare different representation of the text to determine their efficacy in forecasting the success of a firm's campaign. Specifically, we focus on analyzing the business plan. Importantly, our analysis encompasses the entire document rather than isolated words or brief para-

[1]See https://ig.ft.com/generative-ai/ for a synthetic and clear overview.

graphs. We build a novel dataset with more than 640 business plans from seven major Italian crowdfunding platforms, focusing on equity-based crowdfunding.

It is worth highlighting that we employ transformer-based encoder models, distinct from generative models -e.g. ChatGPT-, as they are designed to build a vectorial representation of the input text rather than predicting the next word. Up to now, no paper has used transformer-based encoder models to extract information from complete and unstructured firm documents in the field of alternative finance.

All in all, our research distinguishes itself from prior endeavors through the utilization of a multitude of text mining methodologies, encompassing the latest advancements in the NLP field. Additionally, it employs the entirety of a designated document, namely the business plan, as a comprehensive source for extracting maximal informational content, developing a novel and significant dataset tailored to the domain under investigation.

2 Methods overview: from textual analysis to textual representation

Advancements in data processing and machine learning have opened up novel approaches for analyzing and processing large amounts of textual data. In the following section, we will explore and leverage various methods, ranging from the occurrences frequency, through semi-automatic extraction of relevant textual features, to the state-of-the-art automatic techniques to obtain a compact representation of document.

In particular, we can group methods to represent large documents in three main classes.

1. Basic methods, such as *bag of words* (BoW), represent each document as a vector where each dimension corresponds to a unique word in the vocabulary, and the value of each dimension represents the frequency or count of the corresponding word in the document. Usually, to take into account the fact that some words appear more frequently in general, the counting is substituted with another metrics called TF-IDF (Term Frequency–Inverse Document Frequency). For example, in Cohen, Malloy, and Nguyen (2020), BoW has been used to represent the complete history of quarterly and annual filings by US corporations. Then, these representations are used to construct different measures of similarity, finding that the break in the routine phrasing has strong impact on future firm outcome.

To construct a BoW representation for a document, the text is first tokenized into words, and then a vocabulary of unique words across all documents is created. Each document's vector is then populated by counting the occurrences of each word in the vocabulary within that document. The resulting vectors are often high-dimensional and sparse, with dimensionality equal to the vocabulary size and many zero values. While BoW captures the frequency of words, it disregards the order and context of words in the document, treating each document as an unordered collection of words. Despite its simplicity and lack of semantic information, BoW remains a foundational method for text representation in NLP since many models use it as input to build a more complex text representation.

2. A semi automatic approach to convert text into features involves employing computational techniques to discern and capture semantic or meaningful characteristics embedded within the textual content. Through this approach, key attributes and patterns indicative of the text's underlying semantics are identified and extracted. This extraction is facilitated by leveraging a combination of automated algorithms and human intervention, where automated methods initially sift through the text to detect salient features, such as word frequencies, syntactic structures, and semantic associations. Subsequently, human experts intervene to refine and validate the extracted features, ensuring their relevance and accuracy in encapsulating the semantic essence of the text. This approach strives to achieve a nuanced understanding of the text's meaning, enabling deeper insights and extrapolating specific characteristics of the text such as readability, tone, sentiment (Castellana and Bacciu, 2020; Brunato et al., 2020). Historically, these features measure the complexity of firm's document that has been measured in only a limited context, and yet it is an important and differentiating aspect of the firm (Loughran and McDonald, 2020).

3. The last approach aims construct a succinct representation of the text that encompasses both semantic and syntactic properties. Approaches for encoding word sequences into embedding vectors involve two types of representation, the *word embedding* and *sentence embedding*. The former is constructed using information on local co-occurrence patterns, ensuring that words with similar meanings have proximate vectors. For example, Kaminski and Hopp (2020) show how word and paragraph vector models, applied to text, speech, and video information from crowdfunding projects' descriptions, can enhance the pre-

diction of campaign outcomes compared to standard models based on campaign characteristics.

The latter approach seeks to portray entire sentences or phrases as vectors in a continuous space. These embeddings capture the overall meaning or context of the sentence, considering the interactions between the words and their positions within the sentence.

Sequence models in particular enable the interplay between words to contribute to their meaning. Unlike word embedding models, which assign a static vector to a class, sequence embedding models allow the meaning to be influenced by neighboring words (Ash and Hansen, 2023). This results in distinct vectors being assigned for phrases like "she filed suit under class action" and "she graduated top of her class." Transformers process an entire sequence at once — be that a sentence, paragraph or an entire article — analysing all its parts and not just individual words. A sentence embedding is a function that maps a variable-length sequence of words or tokens from a sentence into a fixed-length vector in a continuous vector space. Formally, let S be a sentence consisting of n tokens $\{w_1, w_2, \ldots, w_n\}$, and let f be a function that generates a sentence embedding:

$$f : S \to \mathbb{R}^d$$

where $\mathbb{R}^d$ is a d-dimensional vector space. The function f encapsulates the semantic meaning, syntactic structure, and contextual information of the sentence S in the resulting vector $\mathbf{v}$, such that similar sentences are mapped close to each other in the vector space.

Various methods can be employed to define the function f and generate sentence embeddings. One common approach is to use word embeddings to represent individual tokens and then aggregate these embeddings to form a sentence representation. For instance, averaging or pooling the word embeddings of the tokens in the sentence can produce a simple yet effective sentence embedding.

Alternatively, neural network-based models, such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), or transformer-based architectures, can be employed to learn more complex and informative sentence embeddings. These models utilize their architectures and training objectives to capture the sequential dependencies, semantic relationships, and contextual nuances present in the sentences, generating high-quality embeddings that can capture the semantic meaning and structure of the sentences effectively.

2.1 Methodology

To evaluate the predictive capacity of different textual representations, we employ distinct models tailored to each representation. Each methodology paves the way for an in-depth exploration of the distinctive attributes inherent in each depiction, specifically we execute:

1. the *Latent Dirichlet Allocation* (LDA), a generative probabilistic model widely used for topic modeling to discover topics present in a collection of documents (Blei, Ng, and Jordan, 2003). Let D be the number of documents, K be the number of topics, and V be the vocabulary size. LDA assumes a generative process where, for each document d, a topic distribution θ_d is drawn from a Dirichlet distribution with parameter α, and for each topic k, a word distribution ϕ_k is drawn from a Dirichlet distribution with parameter β. Then, for each word n in document d, a topic assignment $z_{d,n}$ is drawn from θ_d, and a word $w_{d,n}$ is drawn from the word distribution corresponding to $z_{d,n}$. The goal of LDA is to infer the hidden topic structure by finding the posterior distribution of latent variables, typically approximated using variational inference or Gibbs sampling. The Dirichlet priors α and β control the sparsity of document-topic and topic-word distributions, respectively. The output of LDA includes document-topic distributions and topic-word distributions, providing insights into the thematic structure of the documents. It requires a representation of the matrix of occurrences of the total words, denoted BoW;

2. the READ-IT, the first advanced readability assessment tool for what concerns Italian, which combines traditional raw text features with lexical, morpho-syntactic and syntactic information. READ–IT is a classifier utilizing Support Vector Machines with LIBSVM (Chang and Lin, 2001). It constructs a statistical model by employing feature statistics extracted from a training corpus, given a set of features and the training data. This model is subsequently employed to evaluate the readability of unseen documents and sentences (Dell'Orletta, Montemagni, and Venturi, 2011). The model's output variables encompass various metrics, including the document's length measured in sentences and words, a fundamental composite index reflecting the internal structure of the text, a lexical index assessing Vocabulary Composition, Type/Token Ratio, and Lexical Density, as well as a syntactic index gauging the complexity of sentence structures through measures such as the depth of syntactic trees or subordinate clauses.

3. BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art deep learning model introduced by Devlin et al. (2018). At its core, BERT utilizes a multi-layer bidirectional transformer architecture, which consists of multiple encoder layers with self-attention mechanisms. The self-attention mechanism allows BERT to capture long-range dependencies and contextual information by assigning different weights to different words in a sequence based on their importance in relation to other words in the sequence. This approach enables BERT to understand the semantics and relationships between words more comprehensively than unidirectional models. During pre-training, BERT is trained on a large corpus to predict masked words in sentences, utilizing both left and right contexts. This pre-training enables BERT to learn general language representations that capture syntactic and semantic information. Fine-tuning BERT on specific downstream tasks, such as question answering, sentiment analysis, or named entity recognition, further enhances its performance by adapting its pre-trained representations to the task at hand. BERT's ability to generate rich and contextualized embeddings has led to its widespread adoption and benchmark performance across various natural language processing tasks.

   We implement the italian version, UmBERTo-Commoncrawl-Cased [2] utilizes the Italian subcorpus of OSCAR [3] as training set of the language model. We used deduplicated version of the Italian corpus that consists in 70 GB of plain text data, 210M sentences with 11B words where the sentences have been filtered and shuffled at line level in order to be used for NLP research.

Despite the different text representation that each method take as input, all the methods can be employed to obtain a vector representation of a document d. We use LDA to map a document d into a vector $v \in \mathbb{R}^T$, where T are the number of topics in LDA; the entry v_j indicates the probability to assign the topic j to the document d. In the case of READ-IT, we use the lexical/grammatical statics generated a vector representation u; the entry u_j represents the j-th statics computed by READ-IT on d. BERT is the most

[2]https://github.com/musixmatchresearch/umberto

[3]The OSCAR project (Open Super-large Crawled Aggregated coRpus) is an Open Source project aiming to provide web-based multilingual resources and datasets for Machine Learning (ML) and Artificial Intelligence (AI) applications. The project focuses specifically in providing large quantities of unannotated raw data that is commonly used in the pre-training of large deep learning models.

straightforward method since it naturally maps the input text into a vector e; the entries of e do not have a particular meaning.

2.2 Classification

Give a vector representation x of a document, we employ a Logistic Regression to predict the success of a campaign:

$$P(Succ = 1|x) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}), \tag{1}$$

where $\{\beta_0, \boldsymbol{\beta}\}$ are the parameters that should be learned. In the following, we make explicit the logistic regression model for each vector representation considered (for the sake of simplicity, we have assumed that LDA and BERT generates vectors of size 10^4):

$$\begin{aligned}\boldsymbol{x\beta_{Read}} = \beta_1 Base_i + \beta_2 Lexical_i + \beta_3 Syntax_i \\ + \beta_4 Global_i + \beta_5 NumberWords_i + \beta_6 NumberSent_i,\end{aligned}$$

$$\begin{aligned}\boldsymbol{x\beta_{LDA}} = \beta_1 Topic1_i + \beta_2 Topic2_i + \beta_3 Topic3_i + \beta_4 Topic4_i + \beta_5 Topic5_i \\ + \beta_6 Topic6_i + +\beta_7 Topic7_i + \beta_8 Topic8_i + \beta_9 Topic9_i + \beta_{10} Topic10_i,\end{aligned}$$

$$\begin{aligned}\boldsymbol{x\beta_{BERT}} = \beta_1 Bert1_i + \beta_2 Bert2_i + \beta_3 Bert3_i + \beta_4 Bert4_i + \beta_5 Bert5_i \\ + \beta_6 Bert6_i + +\beta_7 Bert7_i + \beta_8 Bert8_i + \beta_9 Bert9_i + \beta_{10} Bert10_i.\end{aligned}$$

We also implement a baseline which is not based on text: instead it focuses on variable which describe the campaign and the creator. Again, we employ logistic regression to predict the success:

$$\begin{aligned}\boldsymbol{x\beta_{Base}} = \beta_1 Goal_i + \beta_2 RatioGoal_i + \beta_3 MinInv_i \\ + \beta_4 ShareEquity_i + \beta_5 Age_i + \beta_6 StartUp_i + \beta_7 Income_i.\end{aligned}$$

[4]Due to the size of our dataset, we limited the assignment of LDA model probabilities to only 10 classes, and conducted principal component analysis (PCA) on the vector representations derived from BERT.

3 Dataset

We construct an augmented dataset by relying on seven different Italian crowdfunding platforms to retrieve firms' Business plans (*Startups, Crowdfundme, Ecomill, Opstart, WeAreStarting, MamaCrowd and BacktoWork*) and AIDA *Bureau Van Dijk* to derive other financial variables and firm characteristics. Specifically, from each platform, we identify the equity offerings that make up our sample, covering platforms since crowdfunding inception in 2014 to the end of January 2024. In order to accomplish this, we gathered data by web scraping all available documents and campaign data across all platforms. Starting from the population of 694 successfully funded offerings, we exclude 14 mini-bond offerings and 6 real estate. The sample is made of 674 initial equity offerings. For 29 of them, BPs could not be used due to the extension of the file or the language. The final sample is composed of 645 observation.

3.1 Preprocessing

The first major challenge is to clean the document of any non-textual characters without changing the meaning of individual sentences. First, we download the relevant documents from the different platforms, then we extract the text from the main document (the business plan) by means of specific tools (PDFminer). We now execute the pipeline in order to obtain the individual tokens, by leveraging Python package Stanza (Qi et al., 2020) and spaCy (Honnibal and Montani, 2017) to build our own tokenizer to transform the raw text into the three representations described above.

3.2 Text Visualization

Figure 1 shows a visualization of the text vector representations in t-Distributed Stochastic Neighbor Embedding (t-SNE; Van der Maaten and Hinton (2008)), a dimensionality reduction method that is well suited for high-dimensional data visualization.

For each representation, each point represent a document: the position of the point is obtained by reducing (thanks to t-SNE) the high-dimensional vector representation to two dimensions. In addition, the color shade of each document indicates the topic assigned by LDA while the full dot (plus symbol) represents unfunded (funded) campaigns.

In each plot, there are no clusters of successes/failures. This highlights that there is no an easy correlation between all the vector representations

and the campaign success. In Figure 2, we show the most frequent words for each LDA topic. In this case, we consider only 12 topics.

Figure 1

In Figure 2, we show the most frequent words for each LDA topic. In this case, we consider only 12 topics.

Figure 2

4 Results

4.1 Parameters Set-Up and Metrics

To give solid results (Hastie et al., 2009), we perform a double Cross-Validation (CV) with 5 folds in the outer CV and 4 folds in the inner one. For each outer fold, we perform a model selection using the inner CV. The hyper-parameters validated are: the number T of LDA topics, the number D of principal components used to reduce the size of BERT embeddings, and the data loss coefficient C in the logistic regression. The values validated are: $T \in [5, 10, 15, 20, 25]$, $D \in [5, 10, 20, 30, 40, 50]$, $C \in [0.1, 1, 5, 10, 100, 1000]$. To select the best configuration, we choose the one with the best validation accuracy average on the 4 inner folds. The generalisation error estimate of each model is given by averaging the test error obtained by the best configuration on each outer fold. The test error is computed by re-training a new model with the best hyper-parameters configuration on the union of the training and validation test. The metrics used to assess the generalisation performances are the accuracy, the balanced accuracy, and the ROC score:

- *accuracy score* is metric used to evaluate the performance of a classification model. It represents the ratio of correctly predicted instances to the total instances in the dataset (James et al., 2013).

- *balanced accuracy score* avoids inflated performance estimates on imbalanced datasets. It is the macro-average of recall scores per class or, equivalently, raw accuracy where each sample is weighted according to the inverse prevalence of its true class. In our binary case, balanced accuracy is equal to the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate).

- *Roc AUC score* is area under receiver operating characteristic (ROC) curve, which visually represents the performance of a binary classifier

Table 1: Accuracy Measure

	Base	**Read-it**	**LDA**	**BERT**
Accuracy (Training)	64.58	60.72	61.07	74.47
	(1.06)	(1.03)	(2.59)	(1.38)
Accuracy (Test)	63.66	58.34	58.50	**70.68**
	(3.24)	(5.22)	(1.33)	(3.81)
Balanced Accuracy (Training)	68.56	58.27	64.07	76.90
	(0.29)	(0.37)	(0.44)	(1.58)
Balanced Accuracy (Test)	66.56	56.64	61.49	**71.74**
	(4.94)	(4.27)	(4.00)	(2.84)
ROC (Training)	72.71	60.44	69.56	84.89
	(0.25)	(0.43)	(1.29)	(1.35)
ROC (Test)	70.22	60.62	68.92	**80.80**
	(3.62)	(7.72)	(5.39)	(2.99)

Note: N° sample in training is 515, instead n° sample in test is 130.

system as the discrimination threshold undergoes variation. The curve is constructed by graphing the true positive rate (TPR, also known as sensitivity) against the false positive rate (FPR) at different threshold settings, showcasing the trade-off between these metrics. FPR is equivalent to one minus the specificity or the true negative rate.

4.2 Results

The results for each method described its accuracy measures are reported in Table 1. For the sake of comparison, in Table 2 we report the results obtained by adding control variables (such as year of campaign, firm sector, geographic area and equity crowdfunding platforms) during the model assessment.

Accuracy measurements always converge in the different models with (without) the addition of control variables. The most suitable metric is the ROC, therefore, moving forward, we present the results of this metric in the testing phase and without control variables. The text readability model, READ-IT, faces challenges in achieving the performance levels of other models, maintaining an accuracy of 60.62%. Conversely, LDA narrowly falls short of the Base model, achieving an accuracy of 68.92% and 70.22% respectively. Nevertheless, BERT demonstrates significantly superior performance, achieving an accuracy of 80.80%.

Table 2: Accuracy Measure with Control Variables

	Base	**Read-it**	**LDA**	**BERT**
Accuracy (Training)	69.85	68.72	74.07	78.11
	(2.25)	(1.03)	(1.59)	(0.27)
Accuracy (Test)	66.03	67.34	72.50	**77.68**
	(2.24)	(5.22)	(1.33)	(2.81)
Balanced Accuracy (Training)	70.84	66.27	72.07	80.19
	(0.70)	(0.37)	(1.44)	(1.08)
Balanced Accuracy (Test)	66.03	65.64	70.49	**78.74**
	(3.67)	(7.27)	(3.00)	(4.84)
ROC (Training)	73.21	72.44	81.56	87.89
	(1.15)	(4.43)	(1.29)	(1.35)
ROC (Test)	70.25	71.62	79.92	**83.00**
	(3.51)	(4.03)	(4.31)	(3.60)

Note: N° sample in training is 515, instead n° sample in test is 130.

By adding control variables, the results show an increase of predictive power. Notably, the generative LDA model stands out, reaching an accuracy of 79.92% compared to its previous 68.92%. Similar findings are observed for the READ-IT model, which attains an accuracy of 71.62%. Again, the BERT representation is the most performing one, rise to 83.00%. Moreover, it is worth to highlight that the models which leverage *bag of word* and *semi automatic embedding* increase their prediction by 15% to 20%.

5 Robustness check

For further confirmation of result obtain, we apply powerful classifier on our data, Support Vector Machines (SVM). SVM is a supervised learning algorithm that aims to find the optimal hyperplane that best separates the data points of different classes in a high-dimensional space (Hearst et al., 1998).

Formally, let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n data points, where $x_i \in \mathbb{R}^d$ represents a feature vector in a d-dimensional space, and let $y = \{y_1, y_2, \ldots, y_n\}$ be the corresponding labels with $y_i \in \{-1, 1\}$ for binary classification. The goal of SVM is to find a hyperplane defined by $\mathbf{w} \cdot \mathbf{x} + b = 0$ that maximizes the margin between the two classes while minimizing the classification error. Here, $\mathbf{w}$ is the weight vector perpendicular to the hyperplane, $\mathbf{x}$ is the input feature vector, and b is the bias term.

Mathematically, the optimization problem for SVM can be formulated as:

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} \xi_i$$

subject to

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n$$

where ξ_i are slack variables that allow for misclassification, C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification error, and $||\mathbf{w}||$ is the Euclidean norm of the weight vector $\mathbf{w}$.

In cases where the data is not linearly separable, SVM can be extended to handle non-linear separation by using kernel functions, which implicitly map the input data into a higher-dimensional space where it becomes linearly separable. The optimization problem for the kernelized SVM can be expressed similarly, but with the use of a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ that computes the dot product in the transformed space.

For our investigation, the linear kernel was selected to facilitate a comparative analysis with the outcomes derived in the preceding section. Additionally, polynomial and radial basis function (RBF) kernels were employed to evaluate the robustness of the results and ascertain whether the underlying relationships exhibited non-linear characteristics across each model. The results are reported in Table 3.

As anticipated, BERT consistently demonstrates superior accuracy across all kernels, escalating from 92.99% attained with the linear kernel to 95.99% with the nonlinear variant. Conversely, the efficacy of text-based models exhibits substantial disparity across different kernels. Notably, performance is subpar with the linear and polynomial kernels, yielding accuracy rate

of 54.43% and 51.61% respectively. However, with the radial basis function (RBF) kernel, notably more promising results are obtained, peaking at 92.61% with the LDA model, indicative of a nonlinear relationship.

Table 3: Accuracy Measure

		Base	**Read-it**	**LDA**	**BERT**
Linear	Accuracy (Training)	69.95	66.72	61.07	96.11
		(1.60)	(3.43)	(5.79)	(3.67)
	Accuracy (Test)	67.03	65.34	61.50	**92.68**
		(3.64)	(4.22)	(4.73)	(2.51)
	Balanced Accuracy (Training)	59.46	54.57	52.27	98.34
		(3.45)	(1.67)	(4.23)	(1.67)
	Balanced Accuracy (Test)	56.36	54.43	51.61	**92.99**
		(2.23)	(6.24)	(2.98)	(3.76)
Polynomial	Accuracy (Training)	74.95	64.72	58.07	99.11
		(1.30)	(10.43)	(15.79)	(0.27)
	Accuracy (Test)	73.03	62.34	57.50	**95.68**
		(4.64)	(10.22)	(12.73)	(1.51)
	Balanced Accuracy (Training)	56.46	51.57	52.27	99.34
		(4.45)	(1.67)	(4.23)	(0.17)
	Balanced Accuracy (Test)	53.36	50.43	53.61	**93.99**
		(5.23)	(1.24)	(4.98)	(3.76)
RBF	Accuracy (Training)	88.95	79.72	97.07	100.00
		(3.60)	(5.43)	1.79)	(0.00)
	Accuracy (Test)	87.03	74.34	95.50	**97.68**
		(3.64)	(4.22)	(0.73)	(1.51)
	Balanced Accuracy (Training)	88.46	79.57	98.27	100.00
		(3.45)	(5.67)	(0.23)	(0.00)
	Balanced Accuracy (Test)	86.36	75.43	92.61	**95.99**
		(4.23)	(6.24)	(1.98)	(3.76)

Note: N° sample in training is 515, instead n° sample in test is 130.

6 Conclusion

There is increasing evidence of the need to introduce new, non-strictly financial variables, to predict or understand economic phenomena as this might provide information not always available when relying on quantitative sources only. However, interpretating qualitative sources is often arduous and we need to leverage new techniques to analyse variables never analysed before.

In this paper, we contribute to this literature by investigating how much the information content of a Business Plans (BPs) may be used to predict the success of a crowdfunding campaign, applying from the simplest *bag of word* to the latest *sentence embedding* based on Trasformers. BPs are crucial for startups, companies, one-person businesses, and even researchers in academia, seeking approval and funding for new ideas.

The challenge lies in using the entire document to extract informative information. Indeed, the increasing complexity and length of information in these types of documents, as highlighted by Cohen, Malloy, and Nguyen (2020), may explain the limited systematic use of entire document observed so far.

We transform the entire text into several representations, following a thorough pre-processing pipeline, to compare their efficacy in forecasting the success of a firm's campaign. We build a novel dataset with more than 640 business plans from seven major Italian crowdfunding platforms.

Our results show that this direction is promising since BERT-based model have higher performance than the basic model, where only the quantitative data is analysed (i.e. goal, ratio goal, equity share, firm's age). Moreover, the fact the recent NLP techinques (such as BERT) outperforms "classical" methods (such LDA and READ-IT) highlights that there is an exciting unexplored area which combines state-of-the-art NLP tools and econometric models. This finding indicates that the text contains a wealth of information that significantly impacts the success or failure of the financing. Therefore, it is prudent to continue exploring this approach to extract and interpret these variables.

This research not only underscores the crucial role of advanced natural language processing techniques in the realm of equity crowdfunding but also provides actionable insights for entrepreneurs, investors, and platform operators alike, facilitating more informed decision-making processes and fostering sustainable growth within the crowdfunding ecosystem.

We are of the opinion that the application of Deep Learning models to entrepreneurship research presents unique opportunities and addresses empirical and theoretical challenges that have, until now, remained inconclusive for various reasons.

References

Ash, Elliott and Stephen Hansen (2023). "Text algorithms in economics". In: *Annual Review of Economics* 15, pp. 659–688.

Athey, Susan and Guido W Imbens (2019). "Machine Learning Methods that Economists Should Know About". In: *Annual Review of Economics* 11, pp. 685–725.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3(Jan), pp. 993–1022.

Brunato, Dominique et al. (2020). "Profiling-ud: a tool for linguistic profiling of texts". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 7145–7151.

Castellana, Daniele and Davide Bacciu (Dec. 2020). "Learning from Non-Binary Constituency Trees via Tensor Decomposition". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. International Committee on Computational Linguistics: Barcelona, Spain (Online), pp. 3899–3910. DOI: 10.18653/v1/2020.coling-main.346. URL: https://aclanthology.org/2020.coling-main.346.

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen (2020). "Lazy Prices". In: *The Journal of Finance* 75(3), pp. 1371–1415.

Dell'Orletta, Felice, Simonetta Montemagni, and Giulia Venturi (July 2011). "READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification". In: *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*. Ed. by Norman Alm. Association for Computational Linguistics: Edinburgh, Scotland, UK, pp. 73–83. URL: https://aclanthology.org/W11-2308.

Delmar, Frédéric and Scott Shane (2003). "Does business planning facilitate the development of new ventures?" In: *Strategic management journal* 24(12), pp. 1165–1185.

Deng, Liang et al. (2022). "A Literature Review and Integrated Framework for the Determinants of Crowdfunding Success". In: *Financial Innovation* 8(1), p. 41.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as data". In: *Journal of Economic Literature* 57(3), pp. 535–574.

Hastie, Trevor et al. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Hearst, Marti A. et al. (1998). "Support vector machines". In: *IEEE Intelligent Systems and their applications* 13(4), pp. 18–28.

Honnibal, Matthew and Ines Montani (2017). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* 7(1), pp. 411–420.

Hormozi, Amir M et al. (2002). "Business plans for new or small businesses: paving the path to success". In: *Management decision* 40(8), pp. 755–763.

James, Gareth et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Kaminski, Jerzy C and Christian Hopp (2020). "Predicting Outcomes in Crowdfunding Campaigns with Textual, Visual, and Linguistic Signals". In: *Small Business Economics* 55, pp. 627–649.

Loughran, Tim and Bill McDonald (2011). "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks". In: *The Journal of Finance* 66(1), pp. 35–65.

Loughran, Tim and Bill McDonald (2016). "Textual Analysis in Accounting and Finance: A Survey". In: *Journal of Accounting Research* 54(4), pp. 1187–1230.

Loughran, Tim and Bill McDonald (2020). "Textual analysis in finance". In: *Annual Review of Financial Economics* 12, pp. 357–375.

McKenzie, David and Dario Sansone (2019). "Predicting Entrepreneurial Success is Hard: Evidence from a Business Plan Competition in Nigeria". In: *Journal of Development Economics* 141, p. 102369.

Qi, Peng et al. (2020). "Stanza: A Python natural language processing toolkit for many human languages". In: *arXiv preprint arXiv:2003.07082*.

Signori, Andrea and Silvio Vismara (2018). "Does success bring success? The post-offering livesof equity-crowdfunded firms". In: *Journal of Corporate Finance* 50, pp. 575–591.

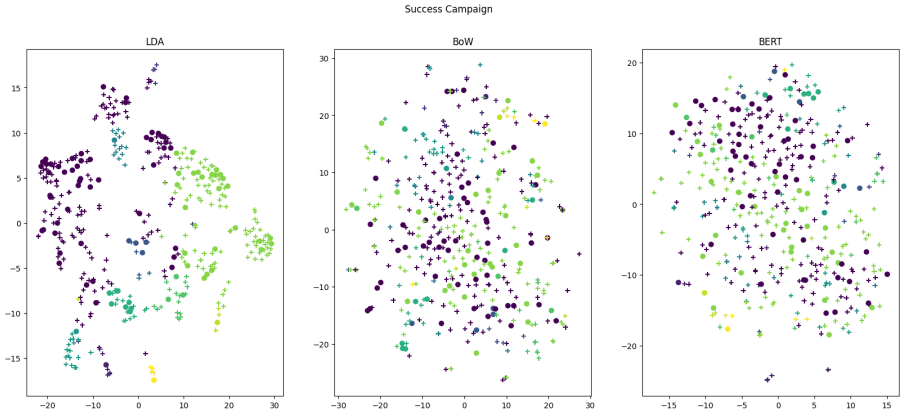
Van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9(11).

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Zhou, Mingming et al. (2018). "Project Description and Crowdfunding Success: An Exploratory Study". In: *Information Systems Frontiers* 20, pp. 259–274.

Ziems, Caleb et al. (2024). "Can large language models transform computational social science?" In: *Computational Linguistics* 50(1), pp. 237–291.

Figure 1: t-SNE on different document vector representations.



Human We employed different classifier with subset of understudy variables.

Figure 2: Word cloud for each topic learned by LDA.

