Annals of Operations Research
https://doi.org/10.1007/s10479-023-05800-w

ORIGINAL RESEARCH

Explainable text-based features in predictive models of crowdfunding campaigns

Viktor Pekar[1] · Marina Candi[2] · Ahmad Beltagui[1] · Nikolaos Stylos[3] · Wei Liu[4]

Received: 1 April 2023 / Accepted: 15 December 2023

Abstract
Reward-Based Crowdfunding offers an opportunity for innovative ventures that would not be supported through traditional financing. A key problem for those seeking funding is understanding which features of a crowdfunding campaign will sway the decisions of a sufficient number of funders. Predictive models of fund-raising campaigns used in combination with Explainable AI methods promise to provide such insights. However, previous work on Explainable AI has largely focused on quantitative structured data. In this study, our aim is to construct explainable models of human decisions based on analysis of natural language text, thus contributing to a fast-growing body of research on the use of Explainable AI for text analytics. We propose a novel method to construct predictions based on text via semantic clustering of sentences, which, compared with traditional methods using individual words and phrases, allows complex meaning contained in the text to be operationalised. Using experimental evaluation, we compare our proposed method to keyword extraction and topic modelling, which have traditionally been used in similar applications. Our results demonstrate that the sentence clustering method produces features with significant predictive power,

Marina Candi, Ahmad Beltagui, Nikolaos Stylos and Wei Liu contributed equally to this work.

✉ Viktor Pekar
v.pekar@aston.ac.uk

Marina Candi
Marina@ru.is

Ahmad Beltagui
a.beltagui@aston.ac.uk

Nikolaos Stylos
n.stylos@bristol.ac.uk

Wei Liu
wei.liu@kcl.ac.uk

1 OIM Department, Aston University, Aston Street, Birmingham B4 7ET, UK

2 Center for Research on Innovation and Entrepreneurship, Reykjavik University, Menntavegur 1, Reykjavik 101, Iceland

3 Innovation and Digitalisation Research Group, University of Bristol Business School, Queens Ave., Bristol BS8 1SD, UK

4 King's College London, Strand, London WC2R 2LS, UK

Published online: 12 January 2024

compared to keyword-based methods and topic models, but which are much easier to interpret for human raters. We furthermore conduct a SHAP analysis of the models incorporating sentence clusters, demonstrating concrete insights into the types of natural language content that influence the outcome of crowdfunding campaigns.

Keywords Predictive modelling · Crowdfunding · Natural Language Processing · Sentence embeddings · SHAP · 3D printing

1 Introduction

Reward-based crowdfunding has emerged as a model for resourcing innovation activities that would otherwise be unlikely to receive support (Belleflamme et al., 2015; Nucciarelli et al., 2017). Due to its novelty, there is limited understanding of success factors, particularly how to craft a crowdfunding campaign that successfully attracts sufficient funding. Crowdfunding is an open call for relatively small contributions from a large number of funders, who contribute financial resources without need for financial intermediaries (Mollick, 2014; Behl et al., 2022). Funders invest on the basis of earning a reward, such as the product or associated benefits (reward-based) or profit-sharing with the founders (equity-based). Individuals seeking resources to produce a product, who we refer to as founders, provide information on the proposed product through a campaign hosted on a crowdfunding platform, such as Kickstarter or Indiegogo. The campaign explains and demonstrates the product, outlines what funding is needed to produce it and details the rewards on offer, including the product itself as well as other benefits such as discounts or additional features (Chakraborty and Swinney, 2021). The immediacy of the interaction can reveal customers' willingness to pay, allows founders to experiment or "fail fast" and can build a business case for further funding through traditional sources (Belleflamme et al., 2015). In the context of reward-based crowdfunding, funders act as customers who decide whether to purchase a product. However, unlike products that are already on the market and can be easily tested, funders must evaluate the promise of a product prior to its production. With no customer feedback to rely on, they must evaluate the information present in the crowdfunding campaign, which is made up of text and images. The content of the text can have an important bearing on the success of campaigns. A crucial challenge, therefore, is ensuring that campaign text communicates the correct messages to ensure funders are convinced to invest in the campaign.

Previous research has identified stylistic characteristics such as the richness and persuasiveness of text (Yeh et al., 2019) and its conciseness (Greenberg et al., 2013) to improve success, while spelling errors or informal language can have the opposite effect (Mollick, 2014). An important but under-researched topic is content analysis that would reveal how founders can signal quality, preparedness and legitimacy, to ensure funders value the product, accept the risk of investing before production and trust the founder to deliver on their promise.

In recent years, the use of Machine Learning (ML) to automate prediction and decision-making in many business applications has increased rapidly. ML algorithms can process a vast number of salient factors that human analysts may struggle to comprehend when making business decisions, and as such, have been widely applied to problems that were previously impossible to predict accurately, such as the assessment of investment risks (Mahbub et al., 2022; Behl et al., 2022), customer segmentation (Mehta et al., 2021), prediction of customer churn (Ahn et al., 2020), dynamic pricing (Ban and Keskin, 2021), personalized

advertising (Choi and Lim, 2020), and assessment of the quality of sales leads (Yan et al., 2015). Despite ML's advantages and potential, a growing concern relates to the lack of transparency of automated, data-driven solutions, which may compromise accountability, fairness and legality (Langer et al., 2021). End users urgently need automated solutions that can explain, in terms that are intuitive for humans, how decisions were reached. Such systems would gain increased trust and increase the value of insight they offer for business problems. To address this need, considerable progress has been made in developing Explainable AI methods, however previous work in this area has largely focused on structured data (Lundberg and Lee, 2017). In this study we focus on explaining human decisions based on interpretation of natural language, thus contributing to a fast-growing body of recent research on the use of Explainable AI in text analytics (see Danilevsky et al. (2020) for an overview).

Predictive features extracted from natural language text are of particular interest in many applications as they can shed light on ideas, beliefs and emotions that influence business decisions. NLP techniques have long been employed to address different practical problems, where they are used to create numerical representations of textual documents, by analysing them in terms of pre-defined lexical categories (Mitra and Gilbert, 2014), sentiment classifications (Desai et al., 2015; Khan et al., 2020), distributed word and paragraph representations (Kaminski and Hopp, 2019), and topic models (Hansen and McMahon, 2016; Thorsrud, 2020; Park et al., 2021), which are subsequently used as input into ML models of various economic phenomena.

To provide a means to study the effect of campaign presentations on their eventual outcome, the present paper proposes a new method, which extracts predictor variables from natural language text based on clustering of sentences by their semantic similarity. Unlike traditional topic models that are constructed over words, our proposed method analyses entire sentences to provide interpretable insights into the contents of documents. In so doing, we offer progress toward interpretable predictive models incorporating evidence from natural language text.

In the experimental evaluation we apply the proposed method to text communications between founders and investors in order to discover textual signals predictive of funding success, and compare it to two methods that have been popular in prior research on crowdfunding success predictions, namely keyword extraction and topic models. Our results show that semantic sentence clusters provide a more transparent and interpretable way to analyse these texts, compared to keywords and topics models. These results suggest that semantic sentence clusters can also be useful for a variety of applications besides predicting the success of crowdfunding campaigns, including market research, customer feedback, and sentiment analysis. Our method has the potential to significantly improve the interpretability of text-based predictive models in many other business applications, such as corporate communications and customer relationship management, and help address some of the concerns regarding their use in decision-making.

Our work makes three important contributions. First, we develop and test a new method to construct predictor models from text based on clustering individual sentences by their meaning. We demonstrate the benefit to interpretability of using text-based analysis on sentences, as opposed to keywords or topics. Second, our method offers automatic labelling of clusters, which represents a substantial improvement over popular topic modelling methods. Finally, we demonstrate the interpretability of predictor features created using our method compared to topic modelling and keyword methods. Theoretically, this paper contributes to improving the explainability and predictive power of text-based open databases emerging from fundraising platforms. From a managerial viewpoint, the current work provides new insights into how analysis of text databases can be improved.

2 Theoretical background

One key tenet of machine learning theory is that it is possible to design a computer program that can learn to predict outcomes of different situations from data and improve its prediction accuracy with experience (Mitchell, 1997). In developing ML models of economic phenomena, researchers have sought to demonstrate that past observations can be used to accurately predict the future development of these phenomena and, more importantly, to explain factors that affect them, as such insights open possibilities to develop policies that can help to achieve desired economic objectives. When it comes to models operating over textual data, machine learning methods are similarly expected to reveal which concepts or ideas expressed within textual documents act as drivers of different economic outcomes.

In this section, we first provide a broad overview of existing research on modelling macro- and micro-economic phenomena with different types of predictive features derived from textual data (Sect. 2.1). We then discuss prior research on using text-based features in models of crowdfunding campaigns, in particular (Sect. 2.1). Finally, we review studies concerned with explainability of models that operate over text-based features (Sect. 2.3).

2.1 Predictive modelling via text for macro and microeconomic phenomena

Scholars have been working on improving predictions for decades. They do so by designing schemes, implementing hypotheses and even attempting to forecast natural events and disasters. Concerning financial markets, the main aspect of making successful predictions relies on recognising, selecting, and measuring the most relevant and critical predictors. Still, some of these predictors may be difficult to identify or measure, as in certain cases they may constantly vary with respect to temporal or spatial parameters.

Nowadays, social media and crowdfunding platforms play a key role in promoting activities, fundraising efforts, and gathering respective data, thus capitalizing on special interests as expressed through online communities (Candi et al., 2018). Thus, the dynamics in utilizing various online platforms represent one of the most striking challenges to the forecasting abilities of private and public institutions worldwide (Elshendy et al., 2018). Already since 2010, text analysis had been reported as a new phenomenon in financial literature, as the availability of electronic financial text had already been on the rise (Cecchini et al., 2010). Text mining is the process of extracting useful information from textual data sources. Apart from simple processing procedures such as eliminating punctuation and capitalization, it describes the process of ultimately converting the text into algorithmically interpretable data, or numerical values (Naderi Semiromi et al., 2020).

Textual data can provide valuable qualitative information. With the recent advances in statistics and computational techniques, researchers are much better equipped to analyse data to serve forecasting efforts (Aprigliano et al., 2023), but more needs to be done to improve prediction accuracy. NLP can be successfully utilized for learning and understanding human language content, thus making it a crucial capability for capturing sentiments. Leveraging NLP techniques to predict financial markets has gradually established the research field of natural language-based financial forecasting and stock market prediction (Xie and Xing, 2013), which can serve as a springboard for applying NLP algorithms in other fields, including crowdsourcing and marketing of novel digital services.

Furthermore, topic modelling (Blei et al., 2003), a widely used NLP method, can find latent topics in text data and then classify the text according to the topics found. For text analysis, there are different proposed approaches and algorithms for topic modelling, such

as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). In addition to improving predictive performance, topic models have been seen as a way to generate readily interpretable units of text analysis; previous research has developed quantitative metrics for evaluating the interpretability of topics such as semantic coherence and topic exclusivity (Park et al., 2021), which could also find application in online funding initiatives.

Moreover, with the advent of machine learning, prediction systems started to draw more attention and different solutions were proposed. New technologies have become game-changers in advancing content-based customer recommendation systems, risk analysis systems for banks, as well as applications for stock markets (Sert et al., 2020). Social media platforms have served as ideal sources for big data, feeding prediction systems (Candi et al., 2018; Stylos and Zwiegelaar, 2019; Pekar, 2020; Stylos et al., 2021). For example, information about cryptocurrencies is spread through various social media outlets and given the benefits of big data, a growing body of the literature is examining the use of text- and news-based measures as sources of information for forecasting and assessing economic events (Park et al., 2021).

From a slightly different perspective, news texts were converted into feature vectors with word representations (Gunduz, 2021). Stock market forecasting models to improve prediction performance by employing unsupervised learning models to analyse investor sentiments with respect to stock prices can take place where explicit labels are not specified when training the network (Baldi, 2012). In contrast to previous studies, Loginova et al. (2021) used well-refined topic-sentiment features, and the added value of the textual features appears to depend on the nature of the text. Moreover, regarding textual data usage for forecasting foreign exchange market developments, Naderi Semiromi et al. (2020) introduce a rich set of text analytics methods to extract information from daily events and propose a novel sentiment dictionary for the foreign exchange market. Using textual data together with technical indicators as inputs to different machine learning models reveals that the accuracy of market pre-dictions depends on the time of release of news/data as well as on text, with features based on term frequency weighting offering the most accurate forecasts (Aprigliano et al., 2023).

Consequently, combining different data sources can lead to more informed price predictions and improve chances of succeeding in funding projects or ventures via digital platforms. Interestingly, until now, the literature has yet to elucidate how features from different data sources affect predictive performance for cryptocurrency prices. Additionally, the area of market prediction, and even more so cryptocurrency, suffers from a lack of high-quality datasets (Loginova et al., 2021). State-of-the-art decision support systems for stock price prediction incorporate pattern-based event detection in text into their projections. Nonetheless, these systems typically fail to account for word meaning, even though word sense disambiguation is crucial for text understanding. In this case, an advanced NLP pipeline for event-based stock price prediction would allow for word sense disambiguation to be incorporated into the event detection process (Hogenboom et al., 2021).

2.2 Approaches and techniques for predicting success of crowdfunding campaigns

Research on crowdfunding was concerned with predicting success vs. failure of a fund-raising campaign, where "success" means whether or not the funding goal was reached by the end of the campaign. These studies explored a variety of indicators of successful campaigns. Company characteristics, such as the number of previous funding applications, funding goal, campaign duration, the social network activity of entrepreneurs, as well as characteristics of campaign presentations, such as the length of the campaign description and the number

of included videos and images, have been shown to have a significant predictive power; classifiers trained over these features achieve accuracies of between 60% and 80% on the success vs failure classification, as reported in various projects (Etter et al., 2013; Mitra and Gilbert, 2014; Du et al., 2015; Lukkarinen et al., 2016; Davies and Giovannetti, 2018; Wolfe et al., 2021). Despite their influence on success, features such as videos are costly to produce, whereas crafting the text of a campaign is more affordable for founders. Evidence suggests words and phrases can inadvertently create unconscious bias against founders (Younkin and Kuppuswamy, 2018) or confidence in their credibility (Peng et al., 2022). Founders must select whether to highlight their own credibility or that of their business idea (Wang et al., 2020) and which characteristics to draw attention to, yet research to date gives limited guidance on how to craft a successful crowdfunding campaign (Lipusch et al., 2020).

More recently, a number of studies have attempted to extract predictive signals from natural language text, i.e. campaign presentations and conversations of entrepreneurs with potential backers, in order to improve the model quality and gain further insights into the factors that increase chances of successful fund-raising (Kang et al., 2020). This work has employed NLP methods similar to those used in previous research on predictive models of other economic phenomena, discussed in the previous section.

A commonly used technique has been the automatic keyword extraction using n-grams, i.e. all possible sequences of words of a predefined length. Mitra and Gilbert (2014) extracted uni-, bi- and trigrams from campaign descriptions, and used them as features in a logistic regression model of success vs failure, alongside "metadata" features such as project duration and project goal. To determine broad semantic groupings of significant predictors, the study used the hand-built Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2007), where words are organised into semantic categories, and found that successful campaigns are characterized by prominence of words relating to cognitive processes, social processes, emotions and senses. Drawing on the psychological theory of persuasion (Cialdini, 2001), it has also been rendered that many significant predictors belong to the semantic fields of reciprocity, scarcity, social proof, social identity, liking, and authority, which correspond to persuasion techniques commonly used in advertising and marketing.

Desai et al. (2015) followed a similar approach, where LIWC and ad-hoc semantic categories were used to extract keywords that describe psycholinguistic properties of texts. Similar to Mitra and Gilbert (2014), the study found that words relating to persuasion techniques are significant predictors of fund-raising success. Training a separate model on the "Rewards" section of campaign descriptions, and analysing its informative features, the researchers obtain a number of interesting insights about the types of rewards that are associated with successful campaigns. Parhankangas and Renko (2017) investigate the hypothesis that certain linguistic styles of communication determine the funding success for social entrepreneurs, but not for commercial start-ups. The authors represented the linguistic style of a text by assigning it scores along several dimensions, such as concrete language, precise language and interactive style; the scores are based on counting occurrences of words from predefined word lists, e.g., the concreteness of the text was measured in terms of the counts of articles, prepositions and quantifiers. Based on this work, the authors make recommendations to social entrepreneurs on the linguistic style to be used in communications with investors. Predictive features in Chaichi (2021) are also keywords extracted using manually compiled word lists, referring to technical characteristics of products. The approach uses a syntactic dependency parser to identify opinion words grammatically linked to product characteristics, which are then used in a model of a campaign to identify characteristics most appealing to backers.

The model developed in Babayoff and Shehory (2022) includes semantic features, such as LIWC keywords, custom crowdfunding-related "buzzword" lists, as well as topic models.

The study finds that a model trained on semantic features alone performs on par with a model containing only metadata features, but the best results are obtained by combining both types of features. The study examines features with a strong correlation with funding success; it includes only LIWC and custom keywords, but not topic models.

To predict project success, Cheng et al. (2019) train a deep neural network model operating on multimodal input: it incorporates features created from text, images and metadata of the project descriptions. Two types of textual features extracted from the full-text of each descriptions are studied: a bag-of-word representation and word embeddings constructed using GloVe (Pennington et al., 2014). In an ablation analysis of the modalities, it was found that textual features are the most useful in the predictive model.

Kaminski and Hopp (2019) experimented with several learning methods, applied to multimodal input. Textual features have been constructed using a Doc2Vec model (Le and Mikolov, 2014), which mapped the text of a document to a fixed-length vector, similar to GloVe word embeddings. Nonetheless, the distributed document representations, while being able to capture complex semantics contained in a document, cannot be used directly to explain which particular semantic properties of the documents have the greatest predictive power. To answer this question, the authors trained a logistic regression model over bag-of-words representations, and examined estimated coefficients on the words. This step reveals, for example, that words relating to the monetary depictions of the fundraisers, such as "money" and "funding", reduce the chances of reaching the funding goal. On the other hand, similar to Mitra and Gilbert (2014), Kaminski and Hopp (2019) find that words relating to excitement ("amazing"), social interaction ("backer", "community", "thank"), and technical inclusiveness ("open source") tend to characterise successful campaigns.

A summary of reviewed papers in presented in Table 1. In a nutshell, previous work has demonstrated that the text of online campaign descriptions contains useful signals about the chances of crowdfunding success; the methods to operationalize these signals have often relied on keyword extraction, custom word lists, and word categories from LIWC. Some studies employed more advanced NLP techniques such as topic modelling and distributed word and document representations, though only a few of them attempted to explain which semantic aspects of a campaign description influence its eventual outcome. These studies used keywords as features within highly explainable logistic regression models. However, interpretation of such models assumes that the meanings of keywords are independent of context, which is clearly an oversimplification, especially in the case of abstract keywords. The accuracy of the interpretation of the keywords based on general-language dictionaries like LIWC depends a lot on their coverage and adaptation to the application domain in question, such as crowdfunding. Another unexplored issue relates to understanding predictive semantic characteristics of campaigns using a broader range of learning methods.

2.3 Explainable AI in text-based predictive models

Machine learning is a subset of AI. In contrast to rule-based AI, where the solution developer designs rules that generate a decision in response to a certain input, machine learning methods derive such rules automatically from a collection of previously observed pairs of inputs and outcomes. For this reason, most machine learning models are seen as "black boxes", which, while capable of a high prediction accuracy, are very difficult for humans to interpret. As our ultimate goal is to reveal factors driving a certain outcome for a crowdfunding campaign, we are primarily interested in applying Explainable AI methods to machine learning models and, specifically, in interpretability of the models' predictions, i.e. processes that can reveal

Table 1 Review of published research on crowdfunding success predictability (source: authors)

References	Prediction problem	Modelling method	Textual features	Quality of best-performing model
Mitra and Gilbert (2014)	success or failure	Penalised logistic regression	Uni-, bi- and trigrams; LIWC categories; custom word lists	2.24% (error rate)
Desai et al. (2015)	success or failure	SVM, Lasso logistic regression, decision trees, Naïve Bayes	LIWC categories, custom word lists, sentiment scores	0.79 (F-score)
Kim et al. (2016)	fundraising ratio (amount raised divided by funding goal)	Ordinary Least Squares	LIWC	0.22 (R2 score)
Parhankangas and Renko (2017)	success or failure	Logistic regression	LIWC; custom word lists	0.37 (R^2 score)
Cheng et al. (2019)	success or failure	Multimodel deep learning network	GloVe word embeddings	0.75 (F-score)
Kaminski and Hopp (2019)	success or failure	Deep neural network, logistic regression, SVM, Gradient Boosting, Multi-Layer Perceptron	Doc2Vec, bag-of-words	0.71 (F-score)
Wolfe et al. (2021)	success or failure; pledged amount	Logistic regression, Ordinary Least Squares	Sentiment score, custom word lists	0.54 (R^2 score)
Chaichi (2021)	success or failure	Lasso logistic regression	Custom word lists	0.69 (accuracy)
Babayoff and Shehory (2022)	success or failure	SVM, J48, Random Forest, LightGBM, SDG, DNN	LIWC; topic models; custom word lists	0.96 (F-score)

to a human user how specific features and their values have been used by a model to generate a prediction.

Model-agnostic methods to explain feature importances, i.e., to quantify the contribution of a feature to a particular prediction, which are not specific to any learning method, have been seen as an attractive possibility to reveal causal relationships between features of an observation and its predicted outcome. Of these, the SHAP algorithm (Lundberg and Lee, 2017) has a number of useful properties, such as intuitiveness and stability of generated explanations (Schlegel et al., 2019; Velmurugan et al., 2021). SHAP is now considered a state-of-the-art post-hoc explainable AI method in predictive modelling, though it is based on an idea originally proposed by Lloyd Shapley in the 1950s. SHAP is effectively different from the classic approach of utilizing Shapley values as it explains every instance of a factor in the data by computing a single marginal contribution for that occurrence. In essence, the SHAP method integrates multiple additive feature importance elements to achieve local accuracy, consistency and value missingness for extracted explanation coefficients.

Previous applications of SHAP to economic problems have normally involved observations represented with a relatively small number of features (e.g., Chew and Zhang 2022; Haag et al., 2022). A number of studies have employed SHAP to study feature behaviour and biases in models with very large numbers of features extracted from textual data. Velampalli et al. (2022) develop several models for sentiment classification of tweets in the context of a targeting marketing campaign. Representing the text of the tweets as well as emojis in terms of embeddings using pre-trained embedding models such as USE and SBERT, the authors examine the SHAP values for these features to provide an insight of their relative contribution to the eventual sentiment polarity identified in a tweet.

Ayoub et al. (2021) address the problem of detecting misinformation regarding the COVID-19 pandemic. Their classification model is trained on a corpus of textual claims related to COVID-19, represented in terms of DistilBERT embeddings (Sanh et al., 2019). SHAP is then utilised to improve explainability of the proposed model, driving an effort to increase the end user's trust in the model's classifications. They evaluated trust in the predictions using between-subjects experiments on Amazon Mechanical Turk. The outcome was very encouraging in terms of detecting misinformation on COVID-19. However, Ayoub et al. (2021) reported that some experiment participants were confused in making predictions based on single words; this has been seriously considered in the current study as a trigger to improve explainable text-based modelling.

In a platform economy context, Davazdahemami et al. (2023) develop a recommender system that uses a machine learning model trained on features capturing similarities of different products; one type of similarity is created by measuring the distance between distributed Doc2Vec embeddings of textual descriptions of the products. A SHAP analysis is then performed to identify important determinants of link formations. SHAP was thus used to provide insights for product developers about the design principles they can incorporate into their development process to help better match products and their prospective users. In other words, SHAP offers an explanatory perspective as this analysis offers a validation tool for corroborating outputs from analytics and/or marketing studies.

To support tourism marketing campaigns, Gregoriades et al. (2021) train a machine learning model that matches marketing content to consumers. Using product reviews of hotels as input, their approach constructs topic models which are then fed as features into a decision tree classifier, alongside features encoding other cultural and economic information on tourists. The proposed solution thus seeks to improve optimization abilities of campaign contents, by targeting tourists based on SHAP values of features extracted from their own word-of-mouth communications.

3 Text-based predictors of crowdfunding success

As previous research has shown, natural language texts contain signals predictive of various economic and social phenomena, which can be operationalised in statistical models. However, in many applied contexts, it is often impossible to know in advance what these textual signals are – they need to be discovered before they can be incorporated into a predictive model. Therefore, in this paper, our focus is on methods that derive semantic representations from text in an unsupervised manner, i.e., without a reliance on precompiled dictionaries or labelled training data. We first present two approaches that have been popular in previous work, automatic keyword extraction (Sect. 3.1) and Contextual Topic Models (Sect. 3.2). After that, we present a novel unsupervised method to construct highly explainable text-based predictors based on semantic clustering of sentences (Sect. 3.3). Subsequently, these three kinds of predictors will be evaluated within a model of raised crowdfunding investment.

3.1 Keyword extraction

Keywords, which can be individual words or sequences of multiple words, have traditionally been used to represent the essential content of a document. Automatic keyword extraction has been employed in many information management problems, where keywords are used to rank documents by relevance to the user query or to classify documents into thematic categories. Most successful keyword extraction methods are based on a combination of linguistic and statistical evidence: linguistic criteria, such as part-of-speech patterns, select candidate keywords, while statistical measures, such as C-Value, TextRank or TFIDF, identify keywords that have the strongest association to particular documents and thus best describe their contents (for a survey, see Astrakhantsev et al. (2015)).

For the purposes of testing keyword extraction as a method to predict crowdfunding campaign success, we extract candidate keywords with the help of the Rapid Keyword Extractor (RAKE) method (Rose et al., 2010), which has been used in studies of various downstream NLP tasks such as topic modelling (Jeong et al., 2019), text generation (Peng et al., 2018), document classification (Haynes et al., 2022) and construction of large-scale knowledge bases (Sarica et al., 2020). RAKE is based on the observation that useful keywords are often sequences of content words such as nouns and adjectives, and rarely contain any punctuation or function words such as articles and prepositions. RAKE uses a list of "delimiters", consisting of punctuation symbols and function words, to split the text of a document into keywords. An experimental evaluation by Rose et al. (2010) has shown that this procedure produces better-quality candidates than ngrams of different sizes.

RAKE was originally designed to be applied to individual documents, and therefore does not use any background corpus based on which measures of importance of a keyword in a document could be calculated. In our application we do have such a collection of documents. To determine how well a keyword represents the meaning of a document it occurs in, we use a TFIDF weighting scheme, which is one of the most popular measures of keyword relevance in information retrieval (Astrakhantsev et al., 2015); the scheme has also been used to select keywords to represent crowdfunding campaign descriptions (e.g., Desai et al., 2015). The document collection is used to calculate the TFIDF score of each candidate keyword t in document d as follows:

$$TFIDF_{t,d} = log(1 + TF_{t,d}) \cdot \frac{N}{log(1 + DF_t)} \tag{1}$$

where $TF_{t,d}$ is the frequency of t in d, DF_t is the number of all documents in the collection, in which t occurred, and N is the size of the document collection; TF and DF are log-transformed to reduce the effect of outliers.

After extracting candidate keywords, we represent the document collection as a document-by-keyword matrix, to be later fed into the model of crowdfunding investment. To that end, we calculate a mean TFIDF score of each keyword across all documents and select k highest-scoring ones to be used in the matrix, determining k using cross-validation.

3.2 Contextual topic models

Topic models (Blei et al., 2003) is an unsupervised method to discover latent topical groups of words in a document collection. Topic models are conventionally constructed using Latent Dirichlet Allocation (LDA), which represents each document as a probability distribution over topics and each topic as a probability distribution over words. Once topic models have been estimated from training data, they can be applied to test-set documents, producing their representations in terms of topics discovered during training. As topics tend to be composed of semantically related words, prominent topics in a document reflect its semantics. Topic models have been widely used in many NLP applications, such as document classification (Rubin et al., 2012), document clustering (Xie and Xing, 2013), exploration of large document collections (Blei and Lafferty, 2007), aspect-based sentiment analysis (Amplayo et al., 2018), and opinion analysis in social media (Thonet et al., 2017). They have also been used in applications beyond text analysis, where topics constructed from textual documents are used as input into predictive models of problems in diverse areas including healthcare (Lehman et al., 2012; Chen et al., 2016; Chiu et al., 2022), finance (Hansen and McMahon, 2016; Thorsrud, 2020), marketing (Jacobs et al., 2016; Li and Ma, 2020), and management (Bao and Datta, 2014).

In our experiments we include Contextual Topics Models (CTMs), an extension of the original LDA algorithm. Bianchi et al. (2021) proposed CTMs in an attempt to improve on LDA in terms of topic coherence. Coherent topics are composed of words with a clear commonality of meaning, such as "apple, pear, lemon, banana, kiwi," and thus can be more intuitively understood by humans. The original LDA uses bag-of-words (BoW) document representations, i.e., treating every document as an unordered set of words, ignoring grammatical and semantic relationships within sentences. To account for contextual meanings of words, CTMs map documents to contextualised document embeddings, their fixed-length vector representations. The document embeddings are produced using RoBERTa, a language model pre-trained on large amount of text using context-independent word embeddings such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014). The document embeddings are used as input for ProdDLA (Srivastava and Sutton, 2017), a method to learn a neural inference network that maps a BoW document to an approximate distribution over topics. ProdLDA proves not only computationally more efficient in estimating probability distributions than LDA, but also produces more coherent topics. CTMs require the number of topics to be set before training; we determine the best value of this hyperparameter through cross-validation.

3.3 Sentence clusters

We next describe a new method to derive explainable predictor features from the text of a document. Similar to topic models, we start with the assumption that every document addresses

several different topics, each to a different degree. Our method aims to discover topics that are common to multiple documents in a collection and then represent each document as a weighted mixture of these topics. However, unlike topic models, we represent topics in terms of complete sentences, rather than individual words. In doing so, we hope to capture more complex ideas contained in text than can be described with unordered lists of words. The method consists of three main stages. First, each document in the training data is cleaned and tokenised into individual sentences; each sentence is then represented in terms of a sentence embedding, i.e., a numerical feature vector. In the second stage, all sentences in the training data are fed into a clustering algorithm operating on sentence embeddings to create topical clusters of sentences. In the third step, in order to make the meaning of sentence clusters easier for a human user to interpret, each cluster is assigned a descriptive label.

Sentence embeddings The first step constructs sentence embeddings using Universal Sentence Encoder (USE) (Cer et al., 2018). The motivation behind USE is to facilitate transfer learning in a wide range of NLP applications of deep neural networks, removing their dependence on large amounts of training data and computational power. USE achieves a high level of generalisation using multi-task learning: it trains a single neural network on multiple, albeit related NLP problems, such as assessing semantic similarity between sentences, sentence-level sentiment analysis, subjectivity classification of sentences. The neural network is trained over word embeddings, and by learning multiple NLP tasks, it also learns how to represent the semantics of a sentence by contextualising the meaning representations of words in a task-independent manner. After the neural network has been trained, the coding sub-graph of the network can be used to encode any new given sentence into a contextualised embedding vector. The trained encoder can thus be used in a new NLP task to produce an accurate representation of sentences without a large training corpus. Subsequent research has indeed shown that sentence embeddings produced with USE can be used to effectively address diverse NLP applications, such as fake news detection Saikh et al. (2019), aspect-based sentiment analysis (AL-Smadi et al., 2023), stance detection in social media (Rashed et al., 2021), semantic search (Sheth et al., 2021), and document clustering (Pramanik et al., 2023).

Sentence clustering In the next step, all sentences found in the document collection are clustered by their meaning using their USE representations. We use K-Means (Macqueen, 1967), one of the most popular clustering algorithms, well-known for its efficiency. Given a set of objects N represented as attribute vectors and an integer k, the desired number of clusters, K-Means searches for a partition of N into k non-hierarchical clusters that minimises the squared Euclidean distance between cluster members and the centroid of the cluster.

Cluster labelling Automatic labelling of document clusters is a well-known problem in the field of information retrieval, with most popular approaches based on determining words that either have a strong association with a cluster, or that are frequent in the document closest to the cluster centroid (Manning et al., 2008). Recent research has proposed to additionally incorporate a hierarchical lexical resource such as WordNet and word embeddings into the process of selecting words to be used as cluster labels (Poostchi and Piccardi, 2018). In our approach, a label for a cluster is created by (1) calculating the centroid of the cluster by averaging sentence embeddings, (2) extracting all keywords from the sentences of the cluster using RAKE, (3) mapping each keyword to a USE representation, (4) retrieving the three keywords that have the greatest cosine similarity to the centroid of the cluster and using them as the cluster label.

After the clusters have been constructed and labelled, the document collection is represented as a document-by-cluster matrix, where cells encode the number of times sentences belonging to the cluster occurred in the document.

4 Experimental evaluation

We evaluate the three types of analysis described above on the task for predicting the success of Kickstarter crowdfunding campaigns in terms of the amount of money raised. We focus on crowdfunding campaigns that are concerned with 3D printing. On the one hand, this technology is currently in the state of fast growth, with its broad adoption seen across many industry sectors. On the other hand, this narrow technical domain presents itself as a suitable case study, where model interpretation can uncover specific, previously unknown insights.

4.1 Data collection

Experimental data for the study was collected from Kickstarter, one of the oldest and most popular reward-based crowd-funding platforms (Frydrych et al., 2014). The data processing workflow is depicted in Fig. 1. The data was downloaded from the Kickstarter website using a custom crawler based on Selenium, a software package for web browser automation. The crawler was restricted to pages in the 3D Printers category of the website, and to only those campaigns, which had completed, had not been cancelled, and where the identity of the author was verified. Each retrieved page was parsed with an HTML parser and the project title, the pledged amount, as well as the text of the campaign description were extracted and recorded into a database. These steps produced data on 267 campaigns, which ran between September 2014 and June 2021.

4.2 Data preprocessing

The texts extracted from Kickstarter pages were found to contain a lot of noise, i.e., extraneous material unrelated to the description of the campaigns, such as boilerplate text (e.g., elements of website navigation, standard copyright notices, etc.), HTML and Javascript code. The texts were cleaned using the following preprocessing steps that are commonly applied to web pages:

- The texts were tokenized into individual words and sentences using the NLTK package (Bird, 2006).
- Sentences written in a language other than English were automatically detected and removed.
- Sentences shorter than 10 and longer than 40 words were removed, as upon initial data exploration such content often turned out to be boilerplate material.

The size of the preprocessed corpus of documents was 251,376 tokens, its vocabulary containing 15,614 unique words. The mean length of a document was 920.8 words (σ=635.7). The mean number of sentences in a document was 45.5 (σ=31.6). The target variable, pledged amount, was log-transformed to diminish the effect of extreme values. Its distribution in the training data is shown in Fig. 2.

The data was split into the training and test parts: 80% (N=213) was used for training and 20% (N=54) was used to testing. The split was performed using stratified sampling on the discretized "pledged amount" variable, to ensure that the target variable has similar distributions in the training and in the test data.

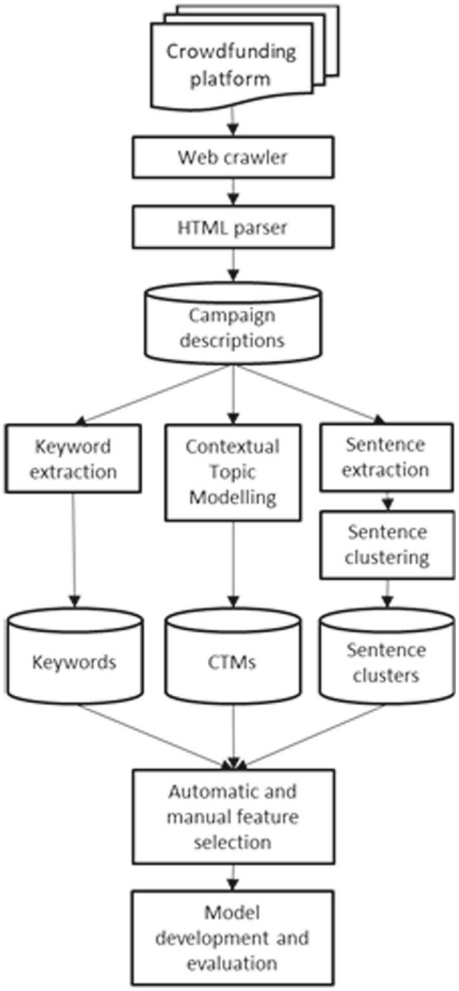


Fig. 1 Data processing workflow

4.3 Feature extraction from text

Keyword extraction Keywords were extracted using the RAKE Python package (Rose et al., 2010). RAKE's delimiter list was extended with additional parts-of-speech—verbs, adverbs, numerals, comparative and superlative adjectives, to give prominence to topical terms that commonly consists of nouns and adjectives. Furthermore, we found that because of part-of-speech tagging errors, RAKE tends to extract word sequences that are too long, but which include useful candidate keywords as substrings. To correct this, we split each candidate proposed by RAKE into bi- and trigrams and recorded them as additional candidates. For example, "strong magnetic coupling mechanism" would be split into "strong magnetic coupling", "magnetic coupling mechanism", "strong magnetic", "magnetic coupling", "coupling mechanism".

Keywords appearing in the title of a document are more likely to attract the reader's attention than those in its main body. To account for this, separate features were created

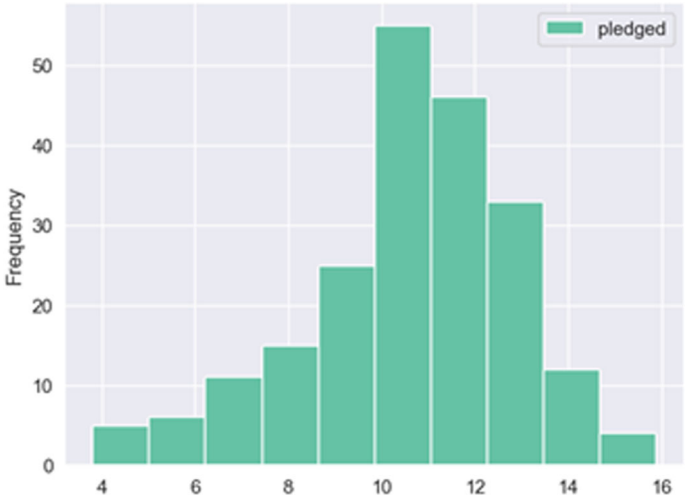


Fig. 2 The distribution of the target variable, "pledged", in the training data

Table 2 Twenty keywords with the largest mean TFIDF weights in the training data

Professional 3d printer_title	make_desc
Professional 3d_title	learn_desc
Laser engraving_title	suggestions_desc
Fdm 3d printer_title	safety_desc
High performance_title	items_desc
High resolution_title	usb cable_desc
Kickstarter fast_title	developed_desc
Zimple_title	connections_desc
Work_title	open source 3d_desc
3d resin_title	laser_desc

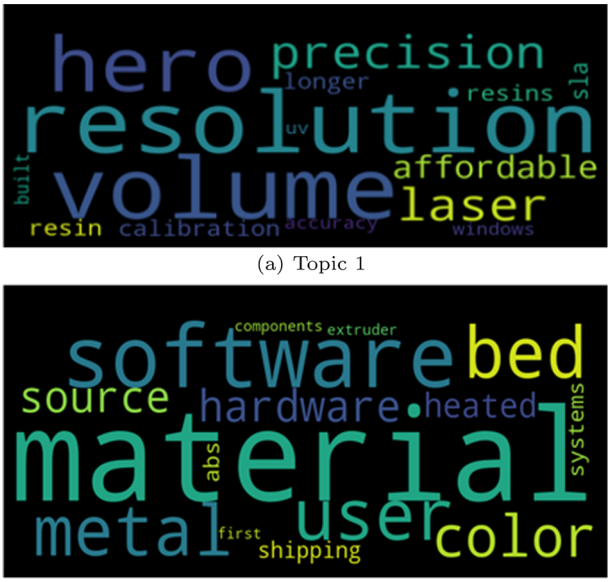
The tags "desc" and "title" indicate if the keyword appeared in the title or body of the document

to indicate if a keyword appeared in the title or in the body of the text by appending a corresponding tag. To purge uninformative keywords, we followed the usual practice and deleted all those that appeared in more than 90% of all documents; additionally, we removed those title keywords with a document frequency of less than three and main-body keywords with a document frequency of less than seven.

Table 2 shows twenty keywords in the training data with the greatest mean TFIDF weights over calculated over training documents.

CTMs To construct CTMs, we use the Python package provided by Bianchi et al. (2021). Because single tokens are used as input into the CTM training process, we performed additional filtering of tokens: the top 100 most common nouns, verbs, and adjectives in English as well as all numerals have been deleted, as their meanings tend to be unrelated to narrow-domain topics such as crowdfunding.

As the pre-trained language model to produce document embeddings, we use Distil-RoBERTa (Sanh et al., 2019), since the original study by Bianchi et al. (2021) recommends the RoBERTa model based on a experimental comparison with alternative models; and we use



(a) Topic 1

(b) Topic 2

Fig. 3 Example topic models. Topic 1 (**a**) contains words relating to precision of 3D printing, while Topic 2 (**b**) to the materials and the printing process. The font size depicts the posterior probability of the words in the topic

its distilled version, due to its compactness and added efficiency. To train a neural inference network predicting topic distribution in a document, we use ProdLDA with 100 epochs.

After training CTMs, the training and the test documents were represented in terms of the discovered topics, to be later used as input for the regression method. Figure 3 provides their examples: the topic in 3a includes words describing the accuracy of printing ("resolution", "precision", "accuracy", "calibration"), while the topic in 3b contains words that appear to relate to the printing process ("material", "metal", "color", "extruder", "bed"). At the same time, it can be noted that the topics contain some words without an obvious pertinence to these topics (e.g., "affordable", "shipping", "first").

Sentence clusters Sentence clusters were created using the USE model pre-trained with a deep averaging network encoder, available on the Tensorflow Hub. Before constructing sentence embeddings, all named entities (names of people, products, companies, etc.) were removed from the sentences, in order to prevent grouping of sentences by their association in the same campaign descriptions. Sentence embeddings were clustered using k-Means, with optimal values of k being determined via cross-validation while training the main regression model. Constructed clusters were labelled with the same keywords that were extracted in the keyword extraction step described earlier in this section. Table 3 shows example clusters created with this process.

The first cluster contains sentences where the emphasis is on low cost of the product, the second cluster—sentences describing professional credentials of the applicant team, and the third cluster—sentences relating to using 3D printers to print edible objects.

Table 3 Examples of automatically labelled sentence clusters

Cluster label: CUT COSTS, AFFORDABLE PRICE TAG, LOWCOST
- In order to ensure high performance and lower costs we made custom electronics
- Decreasing the overall manufacturing cost eventually enables us to offer the all-metal—at a truly affordable price
- Through rigorous design principles, our talented engineering team has been able to cut costs without cutting corners
- Every design choice has been taken in order to reduce the final cost and to increase easy-to-use features, without compromising filament quality
- We may do whatever we can to save the cost, but never in components that will affect the print quality
- To bring the cost down, we combine our technology with an ultra-high resolution panel
- By bringing our product directly to the customer we can keep cost down and innovate quicker
- This way we can offer high quality components but for a far lower price
Cluster label: MAJORING, COMPUTER EXPERT, COMPUTER PROGRAMMER
- He's the tech guru behind the—and was the driving force behind this project's inception
- He has hands-on experience in circuit design and is helping with the electrical aspects of the printer
- That's what led him to become an electronics engineer and a 3D printing expert
- He's had much experience with ergonomics, machine design, product development and other machine-related development
- Apart from working on this project, he is a computer programmer and a talented soccer player
- He spent over a year perfecting the design and in 2015 launched it as a company.
- Leonardo is a computer expert as well as electronic and—is the designer finally—is the marketing expert
- Deeply passionate by the project, he brings his web experience and knowledge to this project
Cluster label: PRINT FOOD, CREATIVE FOOD, FOOD CREATIONS
- You could print an edible—, or a whole pizza, or pasta - all customised you your specific requirements.
- In this way you can personalize your print or food with your own drawing

Table 3 continued

Cluster label: PRINT FOOD, CREATIVE FOOD, FOOD CREATIONS
- Print a holder for your—Pro—, make parts for your model airplane, or print unique and custom baking decorations and or cookie cutters
- Hence you will also have the possibility to print food objects you created yourself
- Create your own design or download a printing template onto your smart device and select the printing conditions for your food capsule
- Within several minutes your individually created and delicious food object will be printed
- What about designing fun shapes on your computer and 3D print them on crackers for an astonishing cocktail?
- Imagine baking at tray of intricate cookies, or cakes printed with the— —

Long dashes stand for deleted named entities (names of people, locations, products, and companies)

Table 4 A comparison of Random Forest and LASSO models trained on text-based features (Keywords, CTMs, and sentence clusters) against a median baseline

	Random Forest				LASSO			
	RMSE	Δ, %	MAE	Δ, %	RMSE	Δ, %	MAE	Δ, %
Baseline	2.33		1.30		2.33		1.30	
Keywords	2.12	−8.85**	1.25	−4.20*	2.33	0.10	1.32	0.88
CTMs	2.09	−10.12*	1.24	−4.87*	2.19	−5.90**	1.27	−2.44*
Sentence clusters	2.22	−4.75	1.29	−1.05	2.18	−6.23*	1.28	−1.59

Statistically significant differences are indicated by asterisks: "*"—α=0.05, "**"—α=0.01 and "***" —α=0.001

4.4 Model evaluation

To create models of pledged investment, we apply two regression algorithms, a polynomial Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1994) and Random Forest (Breiman, 2001). Unlike other widely used learning methods, both algorithms are known to be capable of handling large amounts of features relative to the number of observations, which is the case with our dataset. Each model is trained using text-based features as input and the amount of pledged investment at the end of the campaign as output.

During model training, we determine the best settings for the hyperparameters of the methods (alpha in LASSO, the number of trees, maximum tree depth, the size of the feature subset, the size of training instances subset in Random Forest) using an exhaustive grid search with a ten-fold cross-validation. Once a model with optimal hyperparameter settings was trained, it was applied to the test set to quantify its quality. Considering that the dataset is relatively small, to make use of all data for model evaluation, we performed a test-set evaluation of the best hyperparameter combination using cross-validation with ten folds, i.e. the entire dataset was split into ten parts, a new model was trained on nine parts and evaluated on the last part; the process was run ten times so that each of the ten parts was used as a test set. The evaluation metrics reported for each model in Sect. 5 are averages over the folds.

The baseline in our experiments is a simple model that always outputs the median of the pledged investment estimated from the training data. As evaluation metrics, we use the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Both measure the differences between the model-predicted and ground-truth values, but RMSE gives greater emphasis to large, albeit rare errors than MAE, and so RMSE and MAE can be compared to detect presence of rare large errors.

5 Results and discussion

5.1 The effect of text-based features

In the first set of experiments, we evaluate the predictive power of text-based features in models of funding raised in Kickstarter campaigns. Random Forest and LASSO models were trained on each of the three types of predictors, and compared to the baseline. To determine of differences between models are statistically significant, we use a two-tailed paired t-test. Table 4 shows the obtained RMSE and MAE rates as well as the differences to the baseline.

Table 5 Human agreement on the manual selection of useful predictors among the keywords, CTMs, and sentence clusters

	PSA	%intersection
Keywords	0.36	7.09
CTMs	0.117	4
Sentence clusters	0.607	34.3

Models trained on each type of text-based features outperform the baseline by reducing the error rates by between 4.75% and 10.1%, except for the LASSO model trained on keywords. The reductions in the error rates compared to the baseline are significant with most of the models. The differences between the three non-baseline models are rather small: for Random Forests, they are never greater than 4.3% in terms of RMSE and 3.1% in terms of MAE. For LASSO, the greatest differences are 6.3% in RMSE and 2.4% in MAE. Only in the case of LASSO, models based on CTMs and sentence clusters were significantly better than the model built over keywords, at α=0.001, and only in terms of MAE.

Based on these results, we conclude that keywords, CTMs, and sentence clusters do capture certain signals about the likelihood of a successful campaign, but we were not able to find reliable evidence for one type of textual features having a greater predictive power than others.

5.2 Interpretation of features by human raters

Next, we look at how much human raters agree on their interpretation of the text-based features, namely whether they select the same features as potentially useful indicators of successful crowdfunding campaigns. Two human raters were presented with the lists of keywords, CTMs, and sentence clusters that have been used to create the corresponding best-performing models described above. The raters were instructed to produce binary judgements, i.e. indicate those elements in each list that, in their opinion, reflected a topic or a concept that may potentially influence an investor's decision to make an investment into a crowdfunding campaign.

Table 5 reports the agreement between the two raters in terms of the Positive Specific Agreement (Fleiss, 1975), which is a measure of interrater agreement specifically suited for those cases where one of the two labels presents the main interest, but is much more infrequent than the other label. Because of this, PSA appears particularly suitable in this experiment, as our primary concern is selection of interpretable features. The table shows also the size of the intersection subset, i.e. the subset of cases where both raters assigned the "selected" label to a keyword, CTM or sentence cluster.

We can see that raters have a much greater agreement on the selection of sentence clusters (PSA=0.6) than on the selection of keywords (PSA=0.36). The agreement on selection of useful CTMs is very low (PSA=0.11). The size of the intersection subset is also the largest with the sentence clusters: of all sentence clusters, the percentage of those which both raters deemed as potentially relevant to an investment decision was 34.3% of the original number of clusters, which several times more than such subsets for keywords (7.09%) and CTMs (4%). Based on these results, we conclude that sentence clusters are interpreted similarly by the two raters, while their understandings of the meaning and the relevance of the keywords and CTMs to the problem at hand may be quite different with different human readers of the campaign descriptions.

Table 6 Mean Average Precision scores obtained by Random Forest and LASSO models trained over different types of text-based predictors

	Random Forest	LASSO
Keywords	0.13	0.20
CTMs	0.17	0.20
Sentence clusters	0.17	0.27

Table 7 Models trained on manually selected semantic features compared to corresponding models trained on the full sets of semantic features

	Random Forest				LASSO			
	RMSE	Δ, %	MAE	Δ, %	RMSE	Δ, %	MAE	Δ, %
Keywords	2.38	12.05***	1.32	5.96***	2.34	0.39	1.31	−0.06
CTMs	2.24	7.18***	1.28	3.35*	2.23	1.62	1.28	0.64
Sentence clusters	2.12	−4.34*	1.25	−2.84	2.14	−1.74	1.27	−1.24

Significant differences are indicated with asterisks

In addition, we considered how much the raters' judgements of the features were in agreement with automatic feature ranking. The features were ranked by their SHAP values, as determined within each of the models from the experiments in Sect. 5.1. The quality of each feature ranking was then measured in terms of Mean Average Precision at K (MAP@K), a well-known performance measure used for evaluating document relevance rankings by a Document Retrieval system (Manning et al., 2008). The measure ranges between 0 and 1, with 1 corresponding to the case when all documents manually labelled as relevant to a query appear at the top of the ranked list produced by the system. Table 6 shows the MAP@20 scores obtained for each of the six models.

The feature rankings produced by all the models have low MAP scores (none are greater than 0.27), i.e., human raters did not select many highly predictive features as helping to understand factors affecting a campaign?s success, and the overall interpretability of the models was relatively low. At the same time, we note that models created with sentence clusters have somewhat higher MAP scores than corresponding models with keywords and CTMs, i.e., human-selected sentence clusters tend to be ranked higher by their SHAP values than the other two types of textual features. Thus, this finding further supports our earlier conclusion that it is easier for human raters to identify useful and informative predictors among sentence clusters than among keywords or topic models.

5.3 Manually selected features

To further elaborate the quality of the models, we evaluate the contribution of features that have been selected by both judges, i.e., the interpretable keywords, CTMs, and clusters, within the models of raised crowdfunding capital. Table 7 presents the results of the evaluation.

These results suggest that manual selection helps to improve the sentence clusters models, reducing the error rate by 1.7% to 4.3% compared to models trained on the full set of features. The difference is significant for the RMSE rates of the RF model. On the other hand, we find that keywords and CTMs models do not benefit from manual selection of features, and in fact the models become worse, at a significant level, if trained on human-selected features. The error rates of these sentence clusters are consistently lower than the median baseline:

for Random Forests, the reductions are 8.8% in RMSE (significant at α=0.05) and 3.8% in MAE; for LASSO, they are 7.8% in RMSE (significant at α=0.05) and 2.8% in MAE.

Overall, we conclude that manually selected sentence clusters are not only easier to interpret, but also are better at predicting the success of a campaign than manually selected CTMs and keywords as well as the full set of semantic clusters.

5.4 SHAP feature importances

To obtain insights into which of the manually selected sentence clusters have the greatest influence on the predicted pledged amount of investment, we examined their importances, as determined by the model-independent SHAP (SHapley Additive exPlanations) method (Lundberg and Lee, 2017). To calculate the SHAP values we used the SHAP software package, developed by Lundberg and Lee (2017).

Figures 4 and 5 depict these importances within the Random Forest and the LASSO models. A SHAP value of a feature quantifies its contribution towards the predicted value of the target variable (in this case, the pledged amount of investment for a particular campaign). The features in the plots are sorted by their mean SHAP values in the test set. Thus, the features at the top have a greater influence on the predicted target value than those down the list. Each dot on the plot represents a campaign, with red dots indicating high values of the feature, namely high numbers of sentences present in the document that belong to a particular cluster, while blue dots indicate low values of the features. The position of each dot relative to the horizontal axis shows the SHAP values, with positive SHAP values indicating an increase in the predicted pledged amount compared to a base value (i.e., the average of the target variable across all instances), and negative SHAP values indicating a decrease in the predicted pledged amount.

The ranking of the features in both plots exhibit a lot of similarities, thus confirming conclusions that can be drawn from the plots. The sentence clusters are understandable to humans with knowledge of the context. For example, SIMPLE DESIGN, SIMPLISTIC DESIGN, PARTS EASY (the cluster at rank 1 in both plots) gathers together sentences that relate to simplicity of design and consequent ease of assembly. Here are examples of sentences belonging to the cluster (the values in parentheses show the cosine similarity of the sentence to the cluster centroid):

- "The— —is an extremely simplistic design with few parts which makes assembly a breeze. (0.777)"
- "We've tried to keep assembly as simple as possible with many of the trickier components such as the— — now being pre assembled. (0.648)"
- "The slider is made of fewer components, integrating many functions in a single part, for equivalent performances and easier maintenance. (0.486)"

The large average SHAP value of this cluster suggests that such messages are likely to appeal to investors since simple and elegant design reflects the development effort and hints at quality of the product.

COMPUTER FREE PRINTING, PRINTING FILES, THINGIVERSE (the cluster ranked 2 in both plots) relates to the interface between the 3D printer at the heart of a campaign and the model files that users may print. Sentences in this cluster mention the presence of data storage and processing to allow files to be printed without being connected to a computer. This feature is likely to appeal to investors as it demonstrates the flexibility and usability of the product.

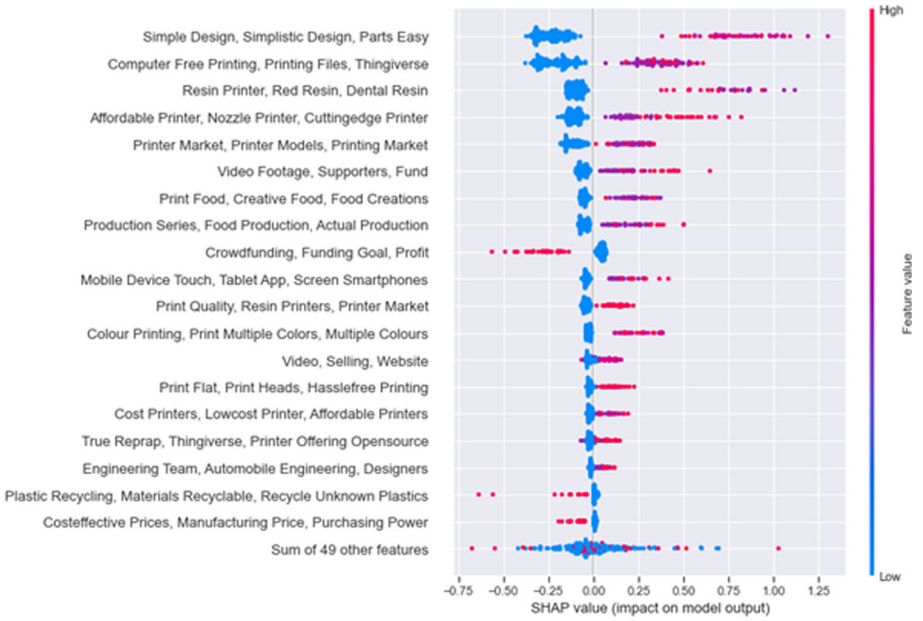


Fig. 4 SHAP values produced using the Random Forest model trained on manually selected sentence clusters

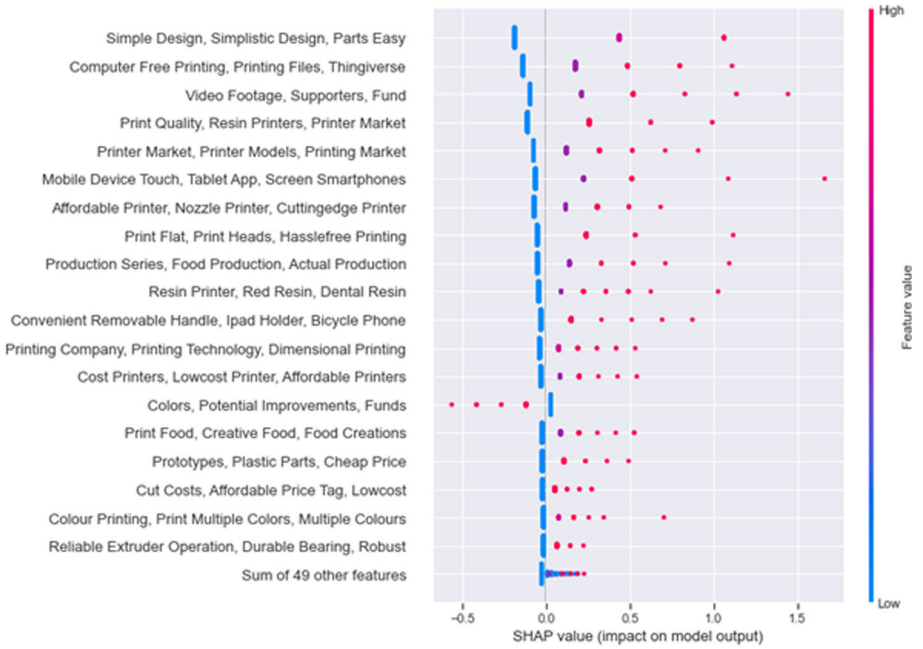


Fig. 5 SHAP values produced using the LASSO model trained on manually selected sentence clusters

Resin Printer, Red Resin, Dental Resin (Cluster 3 in Fig. 4 and Cluster 10 in Fig. 5) relates to resin printers, i.e. those that use a liquid polymer as opposed to the more commonly encountered filament. This cluster is relevant because resin printers for small scale applications and low cost have taken longer to develop than the prevailing filament printers. This represents a new technology that would be attractive to investors and purchasers. These findings tally with the results of a number of previous studies that showed that language used to describe the technical characteristics of a product are predictive of crowdfunding success (e.g., Chaichi, 2021).

The SHAP analysis suggests these sentence clusters to positively relate to investment success, which is intuitive and understandable to humans. The analysis allows ranking of factors, to further support human decision-making. A seemingly counter-intuitive finding, which further demonstrates the value of this analysis are clusters Crowdfunding, Funding Goal, Profit (Cluster 9 in Fig. 4), Cost- Effective Prices, Manufacturing Price, Purchasing Power (Cluster 20 in Fig. 4) and Colors, Potential Improvements, Funds (Cluster 14 in Fig. 5). These clusters mostly relate to achieving funding goals, and the important role that investors or backers have played. For example, Crowdfunding, Funding goal, Profit contains:

- "All we need is your help to reach our goal to produce. (0.638)"
- "We're here because we wouldn't be able to turn our idea into a reality without initial funding. (0.58)"
- "This money isn't going to make us rich or successful, but it is a means of starting on our path. (0.506)"
- "We are a small team with a large vision, and we need your help to make it a reality. (0.466)"
- "We are not dreamers who want to make a one shoot product, we work daily for Fortune 500 companies, we know how to do it. (0.457)"

While these sentences were most likely written with the intention of demonstrating a connection with a community of investors, the SHAP analysis suggests they may be detrimental to success. A plausible explanation is that they cast some doubt in the mind of investors with regard to the risk of achieving the campaign aims. Whereas describing a well-designed product with desirable features and technology suggests a professional approach, suggesting that the campaign can only succeed with the investor's support may have the opposite effect. In particular, reminding the investor of the speculative nature of the vision, or the small size of the team or insisting the campaign is not a dream can have an unexpected negative effect.

It is noteworthy that this finding is very similar to the finding reported in Kaminski and Hopp (2019) that discussion of "legitimizing activities", such as patents, prototypes, or money (as indicated by the significance of corresponding words in BoW representations), are among the worst textual content to be used in a campaign description. This finding is also in line with the observation by Elenchev and Vasilev (2017) that a negative correlation exists between the emphasis on monetary aspects in project descriptions and the outcomes of fundraising campaigns. The SHAP analysis of the models further demonstrates that sentence clusters have the capacity to explain which themes discussed in campaign descriptions attract the attention of potential backers and how they influence the decisions to invest into a project.

6 Discussion and conclusion

6.1 Theoretical implications

This paper addresses the general problem of using text analytics to explain human decisions in business contexts, thus contributing to a growing body of literature on the use of Explainable AI in NLP. The specific application of NLP we investigated was the prediction of reward-based crowdfunding campaign success, a topic, which has recently garnered a lot of attention from researchers due to its immediate practical implications. Using machine learning theory as the starting point, our study attempted to discover textual characteristics that are predictive of successful crowdfunding campaigns from available past observations. The main contributions of our study can be summarized as follows.

The study proposes a new method to extract features from text that can be used in a predictive model of crowdfunding campaigns, potentially alongside non-textual features. The novelty of the method rests on the fact that it is based on semantic clustering of sentences, as opposed to individual words and phrases used in previous research, which enables operationalisation of complex meaning contained in the text. The method is applicable in many other predictive analytics problems, where relevant textual data is available.

Our experimental evaluation has shown that the method is able to produce useful predictive features; the improvement to model quality that the sentence cluster features provide is comparable to the improvement achieved with the established methods of keyword extraction and topic modelling. In addition, experiments with human raters showed that the meaning of sentence cluster features is much easier for humans to interpret than keywords and contextual topic clusters. This points to the ability of the sentence clustering method to provide greater insights into the kinds of information contained in text that are relevant for modelling non-textual phenomena, such as the amount of funds raised in a crowdfunding campaign.

This study offers a key theoretical contribution to operations management theory by shaping explainable models of human decisions based on analysis of natural language text. This type of contribution is exemplified by Holweg et al. (2015) as a key manner for providing key theoretical inputs to operations management theory. Specifically, the fact that this study offers a new way for making better or more inclusive explanations of human decision-making within a crowdfunding campaign context, is an important contribution to the whole natural language processing theoretical body of literature. The fact that this contribution stems from empirical evidence is certainly of great value for operations management literature as per Kilduff et al. (2011).

6.2 Implications for practice

Returning to the original challenge, our research offers limited guidance to founders on how to craft the wording of a campaign in order to improve the chance of success (Lipusch et al., 2020). Where such guidance exists, it might focus on the writing style or emphasis (Wang et al., 2020; Peng et al., 2022), but not necessarily which features or characteristics will appeal most to funders. The method demonstrated in this study can be applied in specific contexts to identify the most pertinent aspects to highlight or avoid while writing the campaign. In the sample of campaigns analysed, text suggesting the campaign depends upon the funder seems to have a negative relation with success. This is in line with studies suggesting founders should avoid emphasising risk or relying too much on emotional appeal over credibility (Majumdar and Bose, 2018). Psychological ownership is viewed positively in crowdfunding

(Nesij Huvaj et al., 2023). Yet in this case there may be a fine line between making potential funders feel they are part of the campaign or leaving them suspicious of whether it can succeed without them. Such subtle differences are difficult to identify or predict, so the method offers important opportunities to analyse the context and select appropriate wording.

In a broader perspective our work also offers guidance to practitioners wishing to base predictions on natural language analysis; they are advised to go beyond traditional keyword-based or topic modelling methods and, instead, use the sentence-clustering method we propose. In doing so, they will achieve increased transparency and improved explainability, both of which will render their predictions more credible as well as more open to scrutiny.

6.3 Limitations and suggestions for future research

There are a number of limitations of this study that future research will hopefully be able to overcome. The proposed feature extraction method was evaluated within a model that included only text-based features. However, as discussed in Sect. 2.2, previous research has shown that there are strong non-textual indicators of success of a crowdfunding campaign, such as campaign duration, fund-raising goal, the experience of the applicants, and others. To understand the relative importance of the text-based features, they will need to be incorporated in a model alongside the non-textual indicators. Evaluation of such a comprehensive model will also demonstrate the practical utility of the insights that the text-based features can provide about the potential success of a given campaign. Another limitation of this study is that the method was evaluated on data from a narrow subject domain, 3D printing technology. Future work may aim to establish if the results we have obtained are generalisable to other types of products that are commonly present on crowdfunding platforms.

Funding Not applicable.

Declarations

Conflict of interest Viktor Pekar declares that he has no conflict of interest. Marina Candi declares that she has no conflict of interest. Ahmad Beltagui declares that he has no conflict of interest. Nikolaos Stylos declares that he has no conflict of interest. Wei Liu declares that she has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate Not applicable.

Code availability Not applicable.

References

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A survey on churn analysis in various business domains. *IEEE Access, 8*, 220816–220839. https://doi.org/10.1109/ACCESS.2020.3042657

AL-Smadi, M., Hammad, M. M., Al-Zboon, S. A., AL-Tawalbeh, S., Cambria, E. (2023). Gated recurrent unit with multilingual universal sentence encoder for Arabic aspect-based sentiment analysis. *Knowledge-Based Systems, 261*, 107540. https://doi.org/10.1016/j.knosys.2021.107540

Amplayo, R. K., Lee, S., & Song, M. (2018). Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis. *Information Sciences, 454–455*, 200–215. https://doi.org/10.1016/j.ins.2018.04.079

Aprigliano, V., Emiliozzi, S., Guaitoli, G., Luciani, A., Marcucci, J., & Monteforte, L. (2023). The power of text-based indicators in forecasting Italian economic activity. *International Journal of Forecasting, 39*(2), 791–808. https://doi.org/10.1016/j.ijforecast.2022.02.006

Astrakhantsev, N. A., Fedorenko, D. G., & Turdakov, D. Y. (2015). Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software, 41*(6), 336–349. https://doi.org/10.1134/S036176881506002X

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat covid-19 infodemic using explainable natural language processing models. *Information Processing & Management, 58*(4), 102569. https://doi.org/10.1016/j.ipm.2021.102569

Babayoff, O., & Shehory, O. (2022). The role of semantics in the success of crowdfunding projects. *PLOS ONE, 17*(2), 1–14. https://doi.org/10.1371/journal.pone.0263891

Baldi, P. Autoencoders, unsupervised learning, and deep architectures. In: Guyon, I., Dror, G., Lemaire, V., Taylor, G., Silver, D. (eds.) Proceedings of ICML Workshop on Unsupervised and Transfer Learning. Proceedings of Machine Learning Research, vol. 27, pp. 37–49. PMLR, Bellevue, Washington, USA (2012). https://proceedings.mlr.press/v27/baldi12a.html

Ban, G.-Y., & Keskin, N. B. (2021). Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science, 67*(9), 5549–5568. https://doi.org/10.1287/mnsc.2020.3680

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science, 60*(6), 1371–1391. https://doi.org/10.1287/mnsc.2014.1930

Behl, A., Dutta, P., Luo, Z., & Sheorey, P. (2022). Enabling artificial intelligence on a donation-based crowdfunding platform: A theoretical approach. *Annals of Operations Research, 319*(1), 761–789. https://doi.org/10.1007/s10479-020-03906-z

Belleflamme, P., Omrani, N., & Peitz, M. (2015). The economics of crowdfunding platforms. *Information Economics and Policy, 33*, 11–28. https://doi.org/10.1016/j.infoecopol.2015.08.003

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 759–766. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2021.acl-short.96. https://aclanthology.org/2021.acl-short.96

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics, 1*(1), 17–35. https://doi.org/10.1214/07-AOAS114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Candi, M., Roberts, D. L., Marion, T., & Barczak, G. (2018). Social strategy to gain knowledge for innovation. *British Journal of Management, 29*(4), 731–749. https://doi.org/10.1111/1467-8551.12280

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems, 50*(1), 164–175. https://doi.org/10.1016/j.dss.2010.07.012

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 169–174). Association for Computational Linguistics, Brussels, Belgium. https://doi.org/10.18653/v1/D18-2029. https://aclanthology.org/D18-2029

Chaichi, N. (2021). Perceived Value of Technology Product Features by Crowdfunding Backers: The Case of 3D Printing Technology on Kickstarter Platform. https://doi.org/10.15760/etd.7580

Chakraborty, S., & Swinney, R. (2021). Signaling to the crowd: Private quality information and rewards-based crowdfunding. *Manufacturing & Service Operations Management, 23*(1), 155–169. https://doi.org/10.1287/msom.2019.0833

Cheng, C., Tan, F., Hou, X., & Wei, Z. (2019). Success prediction on crowdfunding with multimodal deep learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19* (pp. 2158–2164). International Joint Conferences on Artificial Intelligence Organization, Macao, China. https://doi.org/10.24963/ijcai.2019/299

Chen, J. H., Goldstein, M. K., Asch, S. M., Mackey, L., & Altman, R. B. (2016). Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *Journal of the American Medical Informatics Association, 24*(3), 472–480. https://doi.org/10.1093/jamia/ocw136

Chew, A. W. Z., & Zhang, L. (2022). Data-driven multiscale modelling and analysis of covid-19 spatiotemporal evolution using explainable ai. *Sustainable Cities and Society, 80*, 103772. https://doi.org/10.1016/j.scs.2022.103772

Chiu, C.-C., Wu, C.-M., Chien, T.-N., Kao, L.-J., & Qiu, J. T. (2022). Predicting the mortality of icu patients by topic model with machine-learning techniques. *Healthcare*. https://doi.org/10.3390/healthcare10061087

Choi, J.-A., & Lim, K. (2020). Identifying machine learning techniques for classification of target advertising. *ICT Express, 6*(3), 175–180. https://doi.org/10.1016/j.icte.2020.04.012

Cialdini, R. B. (2001). The science of persuasion. *Scientific American, 284*(2), 76–81.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 447–459). Association for Computational Linguistics. https://aclanthology.org/2020.aacl-main.46

Davazdahemami, B., Kalgotra, P., Zolbanin, H. M., & Delen, D. (2023). A developer-oriented recommender model for the app store: A predictive network analytics approach. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2023.11

Davies, W. E., & Giovannetti, E. (2018). Signalling experience and reciprocity to temper asymmetric information in crowdfunding evidence from 10,000 projects. *Technological Forecasting and Social Change, 133*, 118–131. https://doi.org/10.1016/j.techfore.2018.03.011

Desai, N., Gupta, R., & Truong, K. (2015). Plead or pitch? The role of language in kickstarter project success. http://cs229.stanford.edu/proj2015/239_report.pdf

Du, Q., Fan, W., Qiao, Z., Wang, A. G., Zhang, X., & Zhou, M. (2015). Money talks: A predictive model on crowdfunding success using project description. In *Americas Conference on Information Systems*.

Elenchev, I., & Vasilev, A. (2017). Forecasting the success rate of reward based crowdfunding projects. *Econstor preprints, ZBW - Leibniz Information Centre for Economics*. https://EconPapers.repec.org/RePEc:zbw:esprep:170681

Elshendy, M., Colladon, A. F., Battistoni, E., & Gloor, P. A. (2018). Using four different online media sources to forecast the crude oil price. *Journal of Information Science, 44*(3), 408–421. https://doi.org/10.1177/0165551517698298

Etter, V., Grossglauser, M., & Thiran, P. (2013). Launch hard or go home! predicting the success of kickstarter campaigns. In *Proceedings of the First ACM Conference on Online Social Networks. COSN '13* (pp. 177–182). Association for Computing Machinery. https://doi.org/10.1145/2512938.2512957

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31*, 651–659.

Frydrych, D., Bock, A., Kinder, T., & Koeck, B. (2014). Exploring entrepreneurial legitimacy in reward-based crowdfunding. *Venture Capital: An International Journal of Entrepreneurial Finance, 16*, 247–269. https://doi.org/10.1080/13691066.2014.916512

Greenberg, M. D., Pardo, B., Hariharan, K., & Gerber, E. (2013). Crowdfunding support tools: Predicting success & failure. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems. CHI EA '13* (pp. 1815–1820). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2468356.2468682.

Gregoriades, A., Pampaka, M., Herodotou, H., & Christodoulou, E. (2021). Supporting digital content marketing and messaging through topic modelling and decision trees. *Expert Systems with Applications, 184*, 115546. https://doi.org/10.1016/j.eswa.2021.115546

Gunduz, H. (2021). An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination. *Financial Innovation, 7*(1), 28. https://doi.org/10.1186/s40854-021-00243-3

Haag, F., Hopf, K., Vasconcelos, P. M., Staake, T. (2022). Augmented cross-selling through explainable AI—A case from energy retailing.

Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics, 99*, 114–133. https://doi.org/10.1016/j.jinteco.2015.12.008

Haynes, C., Palomino, M., Stuart, L., Viira, D., Hannon, F., Crossingham, G., & Tantam, K. (2022). Automatic classification of national health service feedback. *Mathematics, 10*, 983. https://doi.org/10.3390/math10060983

Hogenboom, A., Brojba-Micu, A., & Frasincar, F. (2021). The impact of word sense disambiguation on stock price prediction. *Expert Systems with Applications, 184*, 115568. https://doi.org/10.1016/j.eswa.2021.115568

Holweg, M., Boer, H., Schmenner, R., Pagell, M., Kilduff, M., & Voss, C. (2015). Making a meaningful contribution to theory. *International Journal of Operations & Production Management*. https://doi.org/10.1108/IJOPM-03-2015-0119

Jacobs, B. J. D., Donkers, B., & Fok, D. (2016). Model-based purchase predictions for large assortments. *Marketing Science, 35*(3), 389–404. https://doi.org/10.1287/mksc.2016.0985

Jeong, B., Yoon, J., & Lee, J.-M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management, 48*, 280–290. https://doi.org/10.1016/j.ijinfomgt.2017.09.009

Kaminski, J., & Hopp, C. (2019). Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals. *Small Business Economics, 55*, 627–649.

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (nlp) in management research: A literature review. *Journal of Management Analytics, 7*(2), 139–172. https://doi.org/10.1080/23270012.2020.1756939

Khan, W., Malik, U., Ghazanfar, M. A., Azam, M. A., Alyoubi, K. H., & Alfakeeh, A. S. (2020). Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. *Soft Computing, 24*(15), 11019–11043. https://doi.org/10.1007/s00500-019-04347-y

Kilduff, M., Mehra, A., & Dunn, M. B. (2011). From blue sky research to problem solving: A philosophy of science theory of new knowledge production. *Academy of Management Review, 36*(2), 297–317. https://doi.org/10.5465/amr.2009.0164

Kim, P. H., Buffart, M., & Croidieu, G. (2016). Tmi: Signaling credible claims in crowdfunding campaign narratives. *Group & Organization Management, 41*(6), 717–750. https://doi.org/10.1177/1059601116651181

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (xai)? - a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence, 296*, 103473. https://doi.org/10.1016/j.artint.2021.103473

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Xing, E. P., Jebara, T. (Eds.), *Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research* (Vol. 32, pp. 1188–1196). PMLR, Bejing, China. https://proceedings.mlr.press/v32/le14.html

Lehman, L. -w., Long, W., Lee, J., & Mark, R. (2012). Risk stratification of icu patients using topic models inferred from unstructured progress notes. *AMIA ... Annual Symposium proceedings/AMIA Symposium*. AMIA Symposium (pp. 505–511).

Li, H. A., & Ma, L. (2020). Charting the path to purchase using topic models. *Journal of Marketing Research, 57*(6), 1019–1036. https://doi.org/10.1177/0022243720954376

Lipusch, N., Dellermann, D., Bretschneider, U., Ebel, P., & Leimeister, J. M. (2020). Designing for crowdfunding co-creation: How to leverage the potential of backers for product development. *Business & Information Systems Engineering*. https://doi.org/10.1007/s12599-019-00628-w

Loginova, E., Tsang, W. K., Heijningen, G., Kerkhove, L.-P., & Benoit, D. F. (2021). Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data. *Machine Learning*. https://doi.org/10.1007/s10994-021-06095-3

Lukkarinen, A., Teich, J., Wallenius, H., & Wallenius, J. (2016). Success drivers of online equity crowdfunding campaigns. *Decision Support Systems, 87*, 26–38. https://doi.org/10.1016/j.dss.2016.04.006

Lundberg, S. M., & Lee, S. -I. (2017). A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *In 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297).

Mahbub, N., Le, A., & Zhuang, J. (2022). Online crowd-funding strategy: a game-theoretical approach to a kickstarter case study. *Annals of Operations Research, 315*(2), 1019–1036. https://doi.org/10.1007/s10479-020-03857-5

Majumdar, A., & Bose, I. (2018). My words for your pizza: An analysis of persuasive narratives in online crowdfunding. *Information & Management, 55*(6), 781–794. https://doi.org/10.1016/j.im.2018.03.007

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mehta, V., Mehra, R., & Verma, S. S. (2021). A survey on customer segmentation using machine learning algorithms to find prospective clients. In *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* (pp. 1–4). https://doi.org/10.1109/ICRITO51393.2021.9596118

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In: Burges, C.J., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 26. Curran Associates, Inc., Harrah's Lake Tahoe, NV. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Mitchell, T. M. (1997). *Machine learning* (Vol. 1). McGraw-hill.

Mitra, T., & Gilbert, E. (2014). The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '14, pp. 49–61. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2531602.2531656

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing, 29*(1), 1–16. https://doi.org/10.1016/j.jbusvent.2013.06.005

Naderi Semiromi, H., Lessmann, S., & Peters, W. (2020). News will tell: Forecasting foreign exchange rates based on news story events in the economy calendar. *The North American Journal of Economics and Finance, 52*, 101181. https://doi.org/10.1016/j.najef.2020.101181

Nesij Huvaj, M., Darmody, A., & Smith, R. S. (2023). Psychological ownership and disownership in reward-based crowdfunding. *Journal of Business Research, 158*, 113671. https://doi.org/10.1016/j.jbusres.2023.113671

Nucciarelli, A., Li, F., Fernandes, K. J., Goumagias, N., Cabras, I., Devlin, S., Kudenko, D., & Cowling, P. (2017). From value chains to technological platforms: The effects of crowdfunding in the digital game industry. *Journal of Business Research, 78*, 341–352. https://doi.org/10.1016/j.jbusres.2016.12.030

Parhankangas, A., & Renko, M. (2017). Linguistic style and crowdfunding success among social and commercial entrepreneurs. *Journal of Business Venturing, 32*(2), 215–236. https://doi.org/10.1016/j.jbusvent.2016.1

Park, E., Park, J., & Hu, M. (2021). Tourism demand forecasting with online news data mining. *Annals of Tourism Research, 90*, 103273. https://doi.org/10.1016/j.annals.2021.103273

Pekar, V. (2020). Purchase intentions on social media as predictors of consumer spending. In *Proceedings of the 14th International AAAI Conference on Web and Social Media. ICWSM 2020* (pp. 545–556). AAAI Press. https://aaai.org/ojs/index.php/ICWSM/article/view/7322

Peng, N., Ghazvininejad, M., May, J., & Knight, K. (2018). Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling* (pp. 43–49). Association for Computational Linguistics, New Orleans, Louisiana. https://doi.org/10.18653/v1/W18-1505. https://aclanthology.org/W18-1505

Peng, L., Cui, G., Bao, Z., & Liu, S. (2022). Speaking the same language: the power of words in crowdfunding success and failure. *Marketing Letters, 33*, 1–13. https://doi.org/10.1007/s11002-021-09595-3

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count (liwc2007).

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics, Doha, Qatar. https://doi.org/10.3115/v1/D14-1162. https://aclanthology.org/D14-1162

Poostchi, H., & Piccardi, M. (2018). Cluster labeling by word embeddings and WordNet's hypernymy. In *Proceedings of the Australasian Language Technology Association Workshop 2018, Dunedin, New Zealand* (pp. 66–70). https://aclanthology.org/U18-1008

Pramanik, A., Das, A. K., Pelusi, D., & Nayak, J. (2023). An effective fuzzy clustering of crime reports embedded by a universal sentence encoder model. *Mathematics*. https://doi.org/10.3390/math11030611

Rashed, A., Kutlu, M., Darwish, K., Elsayed, T., & Bayrak, C. (2021). Embeddings-based clustering for target specific stances: The case of a polarized turkey. *Proceedings of the International AAAI Conference on Web and Social Media, 15*(1), 537–548. https://doi.org/10.1609/icwsm.v15i1.18082

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). 1. Automatic Keyword Extraction from Individual Documents (pp. 1–20). Wiley. https://doi.org/10.1002/9780470689646.ch1

Rubin, T. N., Chambers, A., Smyth, P., & Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Machine Learning, 88*(1), 157–208. https://doi.org/10.1007/s10994-011-5272-5

Saikh, T., Anand, A., Ekbal, A., & Bhattacharyya, P. (2019). A novel approach towards fake news detection: Deep learning augmented with textual entailment features. In E. Métais, F. Meziane, S. Vadera, V. Sugumaran, & M. Saraee (Eds.), *Natural Language Processing and Information Systems* (pp. 345–358). Cham: Springer.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arXiv:1910.01108

Sarica, S., Luo, J., & Wood, K. L. (2020). Technet: Technology semantic network based on patent data. *Expert Systems with Applications, 142*, 112995. https://doi.org/10.1016/j.eswa.2019.112995

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., & Keim, D. A. (2019). Towards a rigorous evaluation of xai methods on time series. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 4197–4201). https://doi.org/10.1109/ICCVW.2019.00516

Sert, O. C., Şahin, S. D., Özyer, T., & Alhajj, R. (2020). Analysis and prediction in sparse and high dimensional text data: The case of dow jones stock market. *Physica A: Statistical Mechanics and its Applications, 545*, 123752. https://doi.org/10.1016/j.physa.2019.123752

Sheth, D., Gupta, A. R., & D'Mello, L. (2021). Using universal sentence encoder for semantic search of employee data. In *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)* (pp. 1–4). https://doi.org/10.1109/ICCICA52458.2021.9697114

Srivastava, A., & Sutton, C. (2017). Autoencoding variational inference for topic models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=BybtVK9lg

Stylos, N., & Zwiegelaar, J. (2019). In: Sigala, M., Rahimi, R., & Thelwall, M. (Eds.) *Big Data as a Game Changer: How Does It Shape Business Intelligence Within a Tourism and Hospitality Industry Context?* (pp. 163–181). Springer. https://doi.org/10.1007/978-981-13-6339-9_11

Stylos, N., Zwiegelaar, J. B., & Buhalis, D. (2021). Big data empowered agility for dynamic, volatile, and time-sensitive service industries: The case of tourism sector. *International Journal of Contemporary Hospitality Management*.

Thonet, T., Cabanac, G., Boughanem, M., & Pinel-Sauvagnat, K. (2017). Users are known by the company they keep: Topic models for viewpoint discovery in social networks. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. CIKM '17* (pp. 87–96). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3132847.3132897

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics, 38*(2), 393–409. https://doi.org/10.1080/07350015.2018.1506344

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Velampalli, S., Muniyappa, C., & Saxena, A. (2022). Performance evaluation of sentiment analysis on text and emoji data using end-to-end, transfer learning, distributed and explainable ai models **13**(2), 167–172.

Velmurugan, M., Ouyang, C., Moreira, C., & Sindhgatta, R. (2021). Evaluating stability of post-hoc explanations for business process predictions. In H. Hacid, O. Kao, M. Mecella, N. Moha, & H.-Y. Paik (Eds.), *Service-Oriented Computing* (pp. 49–64). Cham: Springer.

Wang, W., Chen, W., Zhu, K., & Wang, H. (2020). Emphasizing the entrepreneur or the idea? the impact of text content emphasis on investment decisions in crowdfunding. *Decision Support Systems, 136*, 113341. https://doi.org/10.1016/j.dss.2020.113341

Wolfe, M. T., Patel, P. C., & Manikas, A. S. (2021). Shock and awe: Loudness and unpredictability in twitter messages and crowdfunding campaign success. *Journal of Innovation and Knowledge, 6*(4), 246–256. https://doi.org/10.1016/j.jik.2021.06.002

Xie, P., & Xing, E. P. (2013). Integrating document clustering and topic modeling. CoRR arXiv:1309.6874.

Xie, P., & Xing, E. P. (2013). Integrating document clustering and topic modeling. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence. UAI'13* (pp. 694–703). AUAI Press, Arlington, Virginia, USA.

Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., Chu, S., & Yang, X. (2015). On machine learning towards predictive sales pipeline analytics. *Proceedings of the AAAI Conference on Artificial Intelligence.* https://doi.org/10.1609/aaai.v29i1.9455

Yeh, T.-L., Chen, T.-Y., & Lee, C.-C. (2019). Investigating the funding success factors affecting reward-based crowdfunding projects. *Innovation, 21*(3), 466–486. https://doi.org/10.1080/14479338.2019.1585191

Younkin, P., & Kuppuswamy, V. (2018). The colorblind crowd? founder race and performance in crowdfunding. *Management Science, 64*(7), 3269–3287. https://doi.org/10.1287/mnsc.2017.2774