DOI: https://doi.org/10.15276/hait.09.2026.40
UDC 004.8:65.012.12

Machine learning and sentiment analysis for intelligent project selection: design and experimental validation

Vitaliy M. Kobets[1)]
ORCID: https://orcid.org/0000-0002-4386-4103; vkobets@kse.org.ua. Scopus Author ID: 56006224700
Grygoriy V. Starchenko[1), 2)]
ORCID: https://orcid.org/0000-0003-2707-1055; g.v.starchenko@gmail.com. Scopus Author ID: 57220892211
Pavlo Ye. Ardanov[3), 4)]
ORCID: https://orcid.org/0000-0003-0229-915X; pavlo.ardanov@gmail.com. Scopus Author ID: 54381559600
[1)] Kherson State University, 27, Universitetska Str. Kherson, 73003, Ukraine
[2)] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
37, Prospect Beresteiskyi. Kyiv, 03056, Ukraine
[3)] Institute of Biomedical Technologies, Open International University of Human Development "Ukraine"
23, Lvivska Str. Kyiv, 03115, Ukraine
[4)] University of Oulu, 1, Pentti Kaiteran katu. Oulu, 90570, Finland

ABSTRACT

Relevance. The rapid growth of crowdfunding campaigns on platforms such as Kickstarter, combined with information asymmetry between founders and investors, creates a need for intelligent decision-support tools that account for investor sentiment rather than relying solely on structured financial indicators. **Purpose**. To design and experimentally validate a decision support workflow for predicting crowdfunding campaign success that combines several classification algorithms with sentiment analysis, using backers' comments as an auxiliary consistency check. **Tasks**. To formalize a hybrid feature set uniting structured project metadata with textual sentiment indicators, train and compare classification algorithms across several feature configurations, assess the impact of sentiment on prediction quality, interpret sentiment differences between successful and failed campaigns, cross-check blurb sentiment against comment sentiment, and forecast outcomes for live campaigns. **Methods**. Kickstarter data was processed through deduplication, outlier removal, logarithmic transformation, and one-hot encoding. Five classification algorithms, Logistic Regression, k-Nearest Neighbors, Decision Tree, Random Forest, and Multilayer Perceptron, were trained in three configurations combining structured features with lexicographic and transformer-based sentiment scores; Pearson and Spearman correlations linked blurb and comment sentiment. **Results**. About 194,000 cleaned projects (2009-2026) were analyzed, together with 1.64 million backer comments on the 2,085 overlapping projects. Random Forest performed best; sentiment features gave no meaningful improvement. Failed campaigns showed systematically higher sentiment scores than successful ones, an "over-promising" effect (a weak association, not a causal relationship). A consistency check across projects found a weak but significant correlation between blurb and comment sentiment. **Scientific novelty**. Sentiment is formalized not as a direct predictor but as an inverse quality signal within a hybrid, explainable decision support workflow that jointly models structured metadata, textual sentiment, and community feedback, using backers' comments as an auxiliary consistency signal. **Practical significance**. The workflow can support crowdfunding platforms and investors by providing automated campaign screening, feature-based interpretation of recommendations, and a composite investment-attractiveness index that combines model predictions, sentiment, and expert assessments. **Conclusions**. The workflow provides accurate and interpretable screening of crowdfunding projects. At the same time, its sentiment component requires further validation using full campaign texts and comment time series, as well as testing on other platforms.

Keywords: Intelligent project selection; machine learning; sentiment analysis; crowdfunding; decision support workflow

***For citation*: Kobets V. M., Starchenko G. V., Ardanov P. Ye. "Machine learning and sentiment analysis for intelligent project selection: design and experimental validation". *Herald of Advanced Information Technology*. 2026; Vol.9 No.4: 653–670. DOI: https://doi.org/10.15276/hait.09.2026.40**

INTRODUCTION

The global digital transformation is radically changing the landscape of financial markets, transforming traditional approaches to investment and the selection of promising projects. Historically, the evaluation of start-ups and investment proposals has been based on static econometric calculations, expert analysis of financial statements, and subjective heuristic assessments. However, in today's digital economy, the speed at which data is generated and the sheer volume of data require new, dynamic methods of information processing. Today, the management of investment processes is increasingly viewed as a complex socio- technical system in which decisions are influenced by numerous dynamic factors and a high level of uncertainty [1].

Digitalization has removed geographical and information barriers, enabling investors to analyze thousands of projects in real time. This has led to a shift in focus from simple quantitative models to complex decision support workflows (DSWs)

© Kobets V.., Starchenko, G.., Ardanov P., 2026

This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/deed.uk)

capable of integrating diverse information flows [1]. The need for rapid, objective filtering of investment proposals requires implementing intelligent systems that minimize the human factor, eliminate cognitive biases, and automate the early stages of project screening. Thus, the development of new DSW at the intersection of NLP and financial analysis is a critically important step towards improving the efficiency of capital allocation in the digital economy.

One key manifestation of the digitalization of the investment sector has been the rapid growth of crowdfunding platforms. Platforms such as Kickstarter and Indiegogo have revolutionized the financing of innovation, providing creators with direct access to millions of micro-investors (backers) without intermediaries. Crowdfunding is becoming a vital mechanism for raising funds in the early stages of a project's life cycle, when developers face a lack of start-up capital and the trust of traditional financial institutions.

The Kickstarter platform has given creators direct access to millions of micro-investors without traditional intermediaries. However, the democratization of investment has created a new problem: potential backers lack the time and expertise to conduct a thorough assessment of the thousands of active campaigns. The success rate of Kickstarter campaigns ranges from 40–60% [2], underscoring the urgent need for intelligent tools for preliminary assessment.

Most traditional forecasting models suffer from one-dimensionality: they either consider purely quantitative parameters (funding target, duration), quantitative and qualitative or categorical parameters (country), or focus exclusively on text analysis to identify investor sentiment. The lack of an integrated approach that combines numerical indicators with qualitative characteristics of the project description significantly limits the accuracy and reliability of existing analytical tools.

The integration of machine learning (ML) and natural language processing (NLP) methods offers vast opportunities to overcome these limitations. Modern intelligent systems can analyze both structured numerical indicators and unstructured textual content.

The design and experimental validation of a decision-support workflow based on machine learning and sentiment analysis is a logical and scientifically sound area of research aimed at overcoming the limitations of traditional investment analysis. The integration of natural language processing (NLP) methods and classification algorithms helps avoid one-sidedness in project evaluation by simultaneously analyzing financial and time-based metrics and qualitative communication patterns.

The paper is structured as follows. Section 2 includes related works. Section 3 includes a description of the datasets and the research methodology. Section 4 discloses the research results. Section 5 presents the discussion. The last section concludes.

RELATED WORKS

Decision support workflow for project selection

The design of intelligent decision support workflow for project selection and evaluation is a complex scientific and practical task situated at the intersection of financial technologies, portfolio management, and artificial intelligence methods. The modern scientific paradigm for project selection is shifting from subjective expert assessments [3] and simple linear models [4] towards complex intelligent systems [5], [6] capable of processing both structured numerical indicators and semi-structured textual data.

Predicting the viability and financial success of crowdfunding campaigns has traditionally relied on 'hard' quantitative metadata. Researchers have demonstrated that factors such as the financial target, campaign duration, and the volume of initial contributions have a critical impact on the outcome [7]. However, recent studies show a shift in focus towards a multifaceted analysis of backers' behavior and the dynamics of interaction between project founders and potential investors [8]. In particular, an analysis of the financial success of crowdsourcing platforms and ICOs (Initial Coin Offerings) has revealed that AI-generated assessments and forecasts significantly influence the behavior and decisions of human experts. An important theoretical foundation here is signal theory, which holds that the quality of a project's presentation, the presence of peer reviews, and the transparency of information about the founders serve as positive signals to investors [7]. At the same time, traditional forecasting models suffer from significant limitations in generalizability due to their focus on individual local markets or on limited data samples from a single platform, which reduces their overall real-time validity [9].

The integration of machine learning methods into decision support workflow (DSW) for project selection and portfolio management enables effective modeling of complex, non-linear dependencies that classical statistical methods

cannot address. Early attempts to automate project selection relied on artificial neural networks (ANNs) to classify risks based on key success factors (CSFs) [10]. Neural network approaches have also been successfully applied to incorporate systems thinking into software planning [11]. However, the main drawback of such models remained their 'black box' nature, which limited managers' confidence in them [9].

The current stage of DSW development is characterized by the implementation of explainable artificial intelligence (XAI) methods, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), integrated with powerful ensemble classification algorithms such as XGBoost, Random Forest, or CatBoost [12], [13]. Furthermore, cutting-edge intelligent system architectures are increasingly combining machine learning (ML) with blockchain technology to ensure the transparency and auditability of the expert evaluation process for R&D projects [14], as well as integrating large language models (LLMs) for the automated peer review and analysis of scientific reports [15]. General trends in project analytics point to the rapidly growing role of deep learning, although the adoption of such tools in day-to-day decision-making is hampered by high computational demands and the difficulty of interpreting results for black boxes [16].

A textual description of a project is a key channel for communicating its value to potential investors. Research confirms that the quality of the text, the absence of spelling mistakes, and the frequency of updates are directly correlated with the likelihood of raising funds [7]. Thanks to advances in natural language processing (NLP), it is now possible to measure the emotional tone (sentiment) of textual content. In particular, sentiment analysis on social media and crowdfunding platforms significantly improves the accuracy of predicting the success of start-up funding compared to models based solely on financial metrics [2].

Analyzing sentiment in text descriptions and backers' reviews enables the identification of hidden patterns in risk perception and trust [8]. Researchers combine classical classification models with NLP methods to extract semantic features from campaign text descriptions. However, most existing approaches to sentiment analysis rely on outdated lexical methods or generalized sentiment models that fail to account for the specific context of innovative and technological projects, leading to distorted classification results.

The application of NLP and sentiment analysis to evaluate textual project descriptions has been investigated in [8]. It has been established that the linguistic characteristics of blurbs are statistically significant predictors. FinBERT transformer models [9] deliver higher accuracy in finance-oriented texts compared to VADER's lexical methods.

Predicting the success of crowdfunding campaigns is an active area of research [17] that identified the key role of structured predictors, including financial target, duration, category, and the presence of video content. Ensemble models, notably Random Forest, consistently demonstrate the highest quality metrics in binary classification tasks [16].

Signal theory [7], [8] explains how crowdfunding project authors communicate the quality of their ideas through observable characteristics. In particular, the presence of a 'staff pick' from the platform's editorial team and a video presentation are signals of the author's preparedness and the seriousness of their intentions [18]. ML-based decision support frameworks are actively being researched as tools for reducing investment risks [19].

ML Sentiment Analysis for Crowdfunding Project Selection

Crowdfunding project selection is a decision-making problem under information asymmetry: founders know more about project quality than potential backers or platform screeners, and much of that information is conveyed through unstructured text, such as project descriptions, blurbs, updates, and risk disclosures [20]. These results imply that project selection should not rely on sentiment alone. Crowdfunding prediction studies in the supplied evidence favor hybrid modeling that combines text and structured indicators [21].

Because manually reading large numbers of campaigns is costly, machine-learning sentiment analysis is attractive as a screening aid. Crowdfunding projects include screening, success prediction, risk interpretation, and investor/platform support rather than fully autonomous selection. Explainable AI (XAI) is increasingly important for trust and adoption in screening workflows. XAI can improve crowdfunding success-prediction workflows by identifying which features positively or negatively affect predictions [22].

In crowdfunding, sentiment is predictive but section-dependent [23] due to the heterogeneity of crowdfunding text. Text features (project descriptions, blurbs, titles, founder narratives) are

useful screening inputs, not just descriptive metadata [24]. Sentiment-rich and linguistic features improve crowdfunding prediction when combined with structured data [25].

Machine-learning sentiment analysis is useful for crowdfunding project selection when embedded in a hybrid, explainable workflow [22]. Machine-learning sentiment analysis can support crowdfunding project selection, but the best-supported approach is not a text-only FinBERT pipeline. The evidence supports a hybrid, section-aware decision-support workflow that combines sentiment and linguistic signals from campaign text with structured project indicators such as target amount, updates, supporter counts, and possibly ESG features [21].

For crowdfunding, FinBERT matters because many campaigns, especially investment-oriented or risk-disclosing ones, contain quasi-financial language about targets, returns, uncertainty, and execution. Crowdfunding combines entrepreneurial storytelling with financial and risk-related discourse, especially in sections that discuss goals, execution, uncertainty, and value [26]. A few-shot or zero-shot LLM approach can also match or exceed a well-fine-tuned FinBERT on financial sentiment tasks [27]. FinBERT is not universally best, but it is a strong benchmark, not a guaranteed winner [28]. FinBERT may transfer particularly well to risk disclosures, funding, and target explanations, investor-oriented campaign sections [26].

Evidence on investor behavior suggests users process structured quantitative signals more easily than raw text analytics, so workflow usability may depend on summarizing NLP outputs into concise indicators [29].

A model may predict campaign success well yet still be hard for investors to use. Evidence indicates investors rely heavily on structured quantitative information because it is easier to process than text and images. Therefore, operational value lies in transforming NLP outputs into simple, explainable signals that complement, rather than replace, tabular indicators [29].

Evidence from crowdfunding shows that textual sentiment and linguistic style are not peripheral; they affect funding outcomes [20]. Positive sentiment in blurbs and detailed descriptions contributes to campaign success, and models that use sentiment, linguistic, or persuasion-oriented features outperform simpler TF-IDF baselines [25]. The implication is especially strong for high-volume screening environments where manual reading is impractical [22].

A less ambiguous tone is associated with higher funding success [20]. Positive sentiment in blurbs and detailed descriptions helps success [30]. Models that use linguistic or sentiment features, or persuasion-strategy features, outperform TF-IDF baselines [31]. For workflow design, this implies feature engineering should include polarity, ambiguity/clarity, specificity, persuasion strategies, content emphasis, and narrative framing; section-level sentiment [20], [25], [26].

The most defensible decision-support workflow from the supplied evidence is segment campaign text by function, extract sentiment and linguistic features, combine them with structured project variables, use a strong tabular/hybrid ML model, and surface explanations in concise investor-friendly form [21], [22], [29], [30].

This paper **aims** to address the identified gaps by designing and experimentally validating a comprehensive decision support workflow for the intelligent selection of projects. The scientific novelty of the research lies in the development of a hybrid methodological approach that simultaneously processes structured project parameters (financial targets, deadlines) and dynamic unstructured textual signals (the tone of descriptions and comments on social media) using optimized ensemble classifiers. The proposed solution aims to improve the accuracy of classifying promising projects and to increase the transparency of the selection process by integrating explainable artificial intelligence that enables detailed interpretation of each feature's contribution to the system's final recommendation. Experimental evaluation on real-world data provides preliminary evidence of the approach's robustness within the scope of the present study.

The tasks of the paper is to develop and experimentally validate a decision support workflow for predicting the success of crowdfunding projects, combining five machine learning algorithms with sentiment analysis (VADER, FinBERT) and utilizing a corpus of backers' comments for an auxiliary correlation check between description and comment sentiment.

This study address one-dimensionality gaps through three research questions: (RQ1) Does sentiment analysis (VADER, FinBERT) improve the quality of predictions of crowdfunding campaign success? (RQ2) Which ML algorithm provides the optimal trade-off between accuracy and interpretability for DSW? (RQ3) How does the sentiment of backers' comments correlate with the sentiment of the author's blurb, and what does this

reveal about the reliability of sentiment-based textual signals?

DATA SETS AND METHODS

Dataset description

The study is based on two data sources from main dataset: 83 Kickstarter CSV files (2009-2026), merged and deduplicated to 204,907 unique projects (44 attributes). After cleaning and removing outliers, the final dataset contains 194,388 projects. Distribution of the target variable: successful – 116,154 (59.8 %), failed/canceled – 78,234 (40.2 %). 2,042 live projects were set aside for external prediction. Additional dataset of comments: a JSONL file, containing 5,571,285 comments from backers on 6,312 unique projects; the overlap with the main dataset comprises 2,085 projects (all of which were 'successful' – a distinctive feature of the Kickstarter platform). The main characteristics are demonstrated in Table 1. The 83 CSV files correspond to a single monthly snapshot of the public Web Robots Kickstarter dataset, exported on January 12, 2026 (source folder Kickstarter_2026-01-12); deduplication by unique project id is detailed below. Of the 5,571,285 comments, 1,637,732 (29.4%) belong to the 2,085 overlapping projects and were the ones scored with VADER and aggregated to the project level; for each project we computed the mean, standard deviation and positive/negative/neutral shares of the comment compound score, the comment and unique-author counts (median 408 comments, 785.5 on average per project), and the mean of earlier vs. later comments used to derive sentiment momentum.

The 204,907 unique projects break down by platform state as follows: successful – 116,154; failed – 69,226; canceled – 9,008; submitted – 7,051; live – 2,042; started – 1,421; and suspended – 5 (total = 204,907). The binary target was defined only for terminal outcomes: success (target = 1, n = 116,154) and failure, which combines the 'failed' and 'canceled' states (target = 0, n=69,226+9,008=78,234), forming the labeled set of 194,388 projects. The 2,042 'live' projects were held out as the prediction cohort. The remaining 8,477 projects with non-terminal statuses – 'submitted' (7,051), 'started' (1,421), and 'suspended' (5) – were excluded from both training and prediction, since their funding outcome is not yet determined. Accordingly, 116,154 + 78,234 + 2,042 + 8,477 = 204,907, so every record is accounted for.

Pre-processing included: (1) removal of duplicates by project_id; (2) parsing of JSON fields (category, creator, location, and video); (3) conversion of Unix timestamps; (4) filtering of anomalous dates and zero funding targets. To train the model based on the initial conditions, the following leakage features were excluded: pledged, percent_funded, backers_count, state_changed_at (Temporal Boundary Rule).

Table 1. **Characteristics of the Kickstarter dataset**

Parameter	Value	Comment
Projects (total)	204,907	83 CSV files, 2009-2026, deduplicated by ID
Final dataset after cleaning	194,388	83 CSV files, 2009-2026, deduplicated by ID
Successful from final dataset (target=1)	116,154 (59.8 %)	Class balance is acceptable for ML without SMOTE
Failed/canceled from final dataset (target=0)	78,234 (40.2 %)	Includes 'cancelled' alongside 'failed'
Live projects (prediction)	2042	Unlabeled predictions
Bakers' comments	5,571,285	6,312 projects; 2,085 overlapping with CSV
Basic ML features	40	11 numerical + 29 One-Hot Encoding (OHE) (16 categories + 13 countries)
Sentiment features	6	vader_compound vader_pos, vader_neg; finbert_positive, finbert_negative, finbert_neutral

***Source*: developed by the authors**

Exploratory data analysis

The results of the EDA showed that, amongst the projects for which the funding outcome is already known, the breakdown was as follows: 59.8 % were successfully selected and 40.2 % were rejected (Fig. 1).

The breakdown of project success rates by category (Fig. 2) ranges from the most successful, 'Comics' (79 %), to the least successful, 'Journalism' (23 %).

The duration of both successful and unsuccessful projects has a median of 30 days, but the average for failed projects is 36.8 days, compared with 31.1 days for successful ones (Fig. 3).



Fig. 1. **Distribution of the target variable**
***Source*: compiled by the authors**

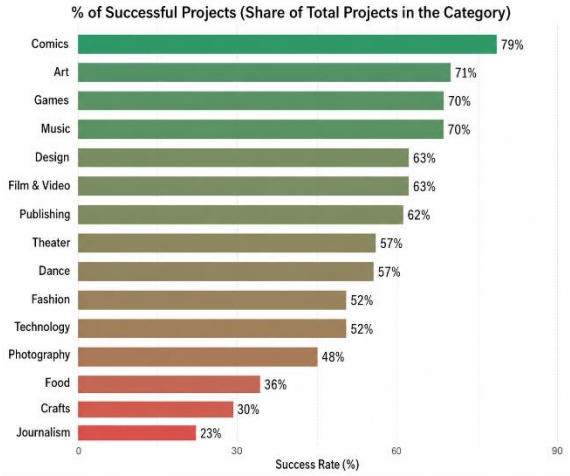


Fig. 2. **Success rate by category**
***Source*: compiled by the authors**

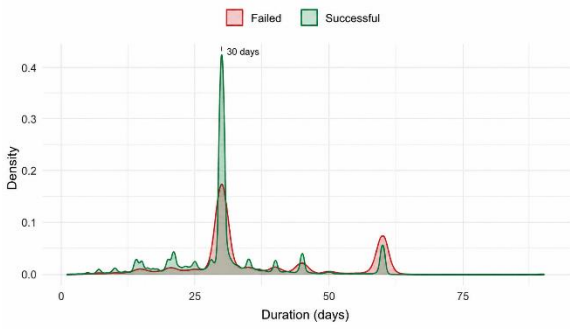


Fig. 3. **Breakdown of campaign duration**
***Source*: compiled by the authors**

Goal (median) is $3500 (successful) vs $8000 (fail). The impact of 'Staff Pick' on the success of start-up funding 91.3 % (yes) vs 54% (no). The impact of a start-up having a video: 66.2 % (yes) vs 46.7 % (no).

The trend in the number of successful campaigns by year, as a proportion of the total number launched, is shown in Fig. 4. Here 2010-2013: success rate ~88 % (small platform, early adopters), a sharp decline in 2014–2015 to ~44–52% (mass exodus), followed by a gradual increase to 68% by 2025.

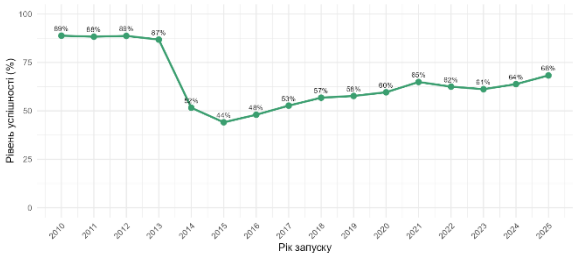


Fig. 4. **Trend in the number of successful campaigns by year (2010-2025)**
***Source*: compiled by the authors**

Hong Kong, HK (83 %) and Japan, JP (74 %) lead the way among countries in terms of the proportion of projects selected (Fig. 5).

The indicators in Table 1 describe the scale, class balance, and feature dimensionality of the empirical base.

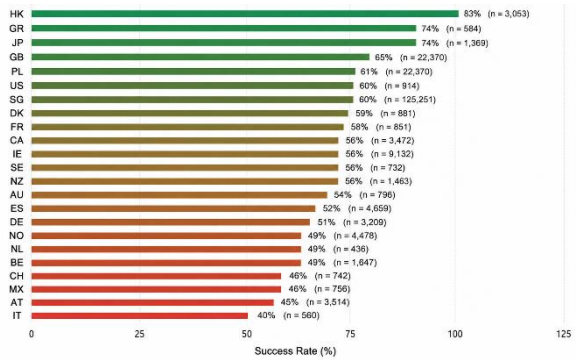


Fig. 5. **Start-up selection success rates by country**
***Source*: compiled by the authors**

The 59.8 % / 40.2 % success-to-failure ratio is computed on the labeled set of 194,388 projects (116,154 successful and 78,234 failed/canceled) and falls within the range that permits training all five classifiers without synthetic rebalancing (e.g., SMOTE), thereby avoiding artificially generated minority-class patterns. The 40 basic predictors (11 numerical and 29 one-hot-encoded: 16 categories and 13 countries) together with 6 sentiment features define the feature space used consistently across all experiments (Tables 2-4). The corpus of 5,571,285 backers' comments, overlapping with main dataset for 2,085 projects, is not used for model training and serves only for the auxiliary correlation analysis in Table 5.

Feature set formation

The feature matrix was constructed using the Temporal Boundary Rule: only features that were

defined at the time the campaign was launched were included. Numeric features: log_goal (log transformation of the financial target), duration_days ([1; 90] days), prep_days ([0;730] days), blurb_length, blurb_words, launch_month, launch_weekday, launch_year, staff_pick (binary), has_video (binary). One-Hot Encoding was applied to 16 categories and 13 countries (threshold ≥0.5 % of the sample).

Funding goals were used in each project's original listing currency without exchange-rate conversion: 36 % of projects are denominated in a non-USD currency (15 distinct currencies across 25 countries; 13 countries are retained as one-hot features at the≥0.5 % threshold). The goal was clipped to [1; 5,000,000] and log-transformed (log_goal = log1p(goal)). Because goals are nominal, the goal medians reported below, are approximate; applying the static USD conversion rate available in the source (static_usd_rate) yields USD-normalized medians of about $3,000 (successful), and $7,200 (failed), versus the nominal $3,500 and $8,000.

Sentiment features were extracted from the 'blurb' field using two methods: (1) lexicographic analysis, where compound VADER [32] ∈ [-1; 1]; values close to +1 indicate the most positive emotional tone of the text, whilst those close to -1 indicate the most negative; values around 0 indicate a neutral tone; (2) FinBERT (ProsusAI/finbert) [33] – a finance-oriented Transformer model, with probabilities of positive/negative/neutral. FinBERT was processed in batches of 32 texts (128 tokens); the total processing time for 204,907 blurbs on a CPU was approximately 6.5 hours. A checkpoint was saved every 2,000 lines.

Machine learning algorithms and tuning

Five algorithms were investigated: Logistic Regression (LR) – a linear baseline, L2 regularisation; K-Nearest Neighbours (KNN, k=21, Euclidean metric) – selection of k via 3-fold cross-validation on a subsample of 20,000; Decision Tree (DT, maxdepth=8, Gini criterion) – pruning based on the optimal complexity parameter (cp); Random Forest (RF, 300 trees, mtry=$\lfloor\sqrt{40}\rfloor$=6) – an ensemble method with feature importance assessment (Gini importance); Multilayer Perceptron (MLP, 64 neurons, decay=0.001, 300 iterations) – a neural network approximation of non-linear relationships.

The implementation was carried out in R 4.5.1 (packages: caret, randomForest, kknn, rpart, nnet, pROC). Split: Train 70 % (136,072) / Test 30% (58,316), drawn by simple random sampling (seed = 2024). The partition was intended to be stratified by the target; however, the target was stored as a numeric 0/1 variable, for which caret::createDataPartition bins the outcome by quantiles, producing a single bin for a binary variable, so no class stratification was actually applied. Consequently, the share of successful projects is 59.29 % in the test set (34,573 of 58,316) and 59.95 % in the training set (81,581 of 136,072), compared with 59.75% in the full labeled set (116,154 of 194,388). A sensitivity check that re-trained the Random Forest with identical settings on a properly stratified split (59.75 % in both partitions) yielded F1=82.94 % and AUC=86.34 %, versus 83.00 % and 86.34 % reported in Table 2, confirming that the small class-share deviation does not affect the results. All metrics in Table 2, Table 3 and Table 4 refer to the original test set. Feature scaling (center + scale) was performed only on the Train set and applied to the Test set (Pipeline Isolation Rule). seed=2024. Each of the five models was trained in three configurations: Baseline (40 features), +VADER (43 features), and +FinBERT (46 features).

Model configuration and selection followed a train/test ratio. Data-dependent transforms did not use test-set information: feature scaling was fitted on the training partition only (Pipeline Isolation Rule), and the data-driven hyperparameters were tuned within the training data – the number of neighbors k for KNN via 3-fold cross-validation on a training subsample, and the complexity parameter cp for the Decision Tree via its internal cross-validation; the remaining hyperparameters (RF: 300 trees, mtry = 6; MLP: 64 neurons, decay = 0.001, 300 iterations) were fixed a priori. The 30 % hold-out was then used to compare the 15 configurations (5 algorithms × 3 feature sets) and to select the final model; accordingly, the reported test metrics of the selected model should be read as a mildly optimistic estimate, since no separate validation partition was reserved for selection. 70/30 split is random rather than temporal: it evaluates in-distribution performance over 2009-2026 but does not, by itself, establish transfer to future campaigns; a temporal hold-out (training on earlier launch years and testing on the most recent ones) is left for future work as the more appropriate protocol for the forward-looking scenario.

Evaluation metrics

Let us consider Accuracy = (TP + TN)/N; Precision = TP/(TP + FP); Recall = TP/(TP + FN); F1 = 2·P·R/(P + R); ROC-AUC; Specificity = TN/(TN + FP); Balanced Accuracy = (Recall +

Specificity)/2. The main criterion is the F1-score (which is robust to class imbalance). Correlation analysis: Pearson's r and Spearman's ρ between sentiment blurb and sentiment of comments (Pearson's test, α=0.05).

RESULTS

Decision support workflows (DSW) are interactive computer systems designed to assist decision-makers in solving weakly structured and unstructured problems using analytical models and databases [14]. In the context of project management and investment analysis, the key task of a DSW is to minimize risk when selecting the optimal option from a set of alternative projects. DSW have come a long way, evolving from the simplest systems based on rigid rules and spreadsheets to modern, artificial intelligence-driven (AI-driven DSW) systems.

The main difference between traditional and intelligent DSW lies in their flexibility and learning capabilities. Traditional systems are limited by predefined formulas and expert weights, which make them susceptible to evaluators' subjectivity [15]. In contrast, intelligent DSWs dynamically adjust their internal parameters based on new data, taking into account not only quantitative metrics (budget, duration, geography) but also complex qualitative indicators, such as the founders' social capital, network connections, and the linguistic characteristics of textual descriptions [12].

The phenomenon of information asymmetry is a fundamental characteristic of two-sided digital platforms, where one side (project authors/founders) possesses comprehensive information about the actual quality, viability, and technological risks of an innovative product, whilst the other party (backers/investors) is forced to make decisions under conditions of high uncertainty [7], [8]. In traditional capital markets, this asymmetry is mitigated through institutional intermediaries. However, in crowdfunding, the absence of strict legal regulation and the significant geographic distance between the parties create risks of adverse selection and moral hazard [17].

To overcome these barriers, Signal Theory is actively applied; it describes the process by which the creators of high-quality projects (signalers) convey reliable information about their hidden characteristics to investors (receivers) to differentiate themselves from low-quality counterparts [8]. According to the theory, for a signal to be effective, it must meet two key requirements: it must be associated with real costs for the sender (a 'costly signal'), which makes it unprofitable for low-quality actors to imitate, and it must be easily observable by the receiver [7]. An example of such a signal in commodity markets is advertising expenditure as a sunk cost and guarantees as an indicator of sellers' reputation.

In crowdfunding campaigns, several categories of quality signals are distinguished:

1. Financial and organizational signals–these include the funding target, the campaign duration, the company's length of existence, and the size of its authorized capital. A funding target that is too high may signal that the founders' plans are unrealistic, which reduces the likelihood of success [17].

2. Social capital and networking signals–the number of followers on social media, the founder's centrality or PageRank within their network of interactions, act as a powerful signal of trust. The presence of well-developed social connections significantly reduces information asymmetry, as investors perceive this as a guarantee of the founder's reputational accountability.

3. Textual descriptive cues–narrative style, the length of the project's textual description, the use of specific vocabulary, and the openness of information about the founders–are important indicators of transparency [7]. The textual description (blurb, full description) is the primary channel for communicating the project's value proposition. Investors analyze the narrative both in terms of its cognitive content and through the prism of its affective (emotional) tone, seeking to assess the team's sincerity and competence [8]. The ability to automatically analyze these signals using machine learning algorithms and NLP opens new opportunities for building intelligent decision-support systems.

Machine learning lies at the heart of modern predictive analytics in project management, transforming the decision-making process from subjective expert analysis to a data-driven, objective approach [12]. In the context of project selection, predicting success largely reduces to a binary classification problem, where the target variable $Y \in \{0,1\}$ indicates whether a campaign succeeded (e.g., achieving a financial target on the platform) [2].

To solve this problem, a supervised learning approach is used, in which the algorithm is trained on a historical data sample $D=\{(x_1,y_1),(x_2,y_2)\ ,\ldots,(x_n,y_n)\}$, where x_i is the vector of project features, and yi is the known outcome [2].

Each machine learning algorithm has its own advantages and limitations for decision support workflows. Logistic regression and decision trees

provide high interpretability of results but are inferior to more complex models in their ability to detect non-linear relationships. In contrast, random forests and multilayer perceptrons demonstrate high accuracy and efficiency in modeling complex relationships among features, though they require additional means to explain their results. The most balanced approach for use in decision support workflows is the random forest, which combines high robustness to noise, the ability to handle non-linear data, and the capability to analyze feature importance.

Sentiment analysis is an important natural language processing tool that enables the identification of emotional and cognitive signals in textual descriptions of projects and their use as additional predictors of crowdfunding campaign success. Unlike structured project characteristics, textual data contains information about the level of optimism, persuasiveness, and professionalism in the presentation of an idea, which influences potential investors' trust. Both lexical approaches (notably VADER) and contextual deep learning models (BERT and FinBERT) are used for sentiment analysis, offering different balances between speed, accuracy, and the ability to account for semantic context. The FinBERT model is the most promising for assessing the investment attractiveness of crowdfunding projects, as it specializes in financial and business terminology and delivers the highest-quality analysis of startup texts and investment proposals.

Based on a synthesis of decision support workflow [14], signal theory [7], [8], machine learning [16], and sentiment analysis of text content [9], a comprehensive conceptual model for the intelligent selection of crowdfunding projects has been developed.

The main objective of the model is to reduce information asymmetry between project creators and investors by converting unstructured text and structured metadata into accurate success predictions and calculating an integrated project attractiveness index. The operational logic of the developed conceptual model is presented in Fig.6.

Stage 1 involves merging the initial CSV files from the Kickstarter platform, removing duplicates from the sample, correcting variable types, unpacking JSON structures, and generating basic analytical metrics.

Stage 2 focuses on visualizing the distribution of target variables and assessing the individual impact of key factors (such as the funding target, campaign duration, or the presence of a video) on the project's ultimate success.

Stage 3 involves selecting predictors that are uncorrelated with the outcome to prevent data leakage, filtering out outliers, applying a logarithmic transformation to the total amount raised, and one-hot encoding categorical variables.

In Stage 4, the data is split into training and test sets, five classification algorithms (LR, KNN, DT, RF, MLP) are run and optimized, and the winning model is selected based on the F1-score.

Stage 5 is dedicated to validating the selected models using a confusion matrix and ROC curves, and to calculating the probabilities of success for current active projects.

In Stage 6, the VADER and FinBERT models are used to analyze the emotional tone of short project descriptions (blurbs) and backers' comments, identifying hidden patterns in the behavior of authors and sponsors.

The final stage of the model involves retraining the classifiers by integrating sentiment metrics to assess their impact on prediction accuracy and to provide a theoretical basis for the behavioral signals of start-up founders. In the final stage, the system generates a ranked list of projects and provides a transparent explanation of the assessments obtained using explainable artificial intelligence methods, thereby enhancing the soundness of and confidence in the investment decisions made.

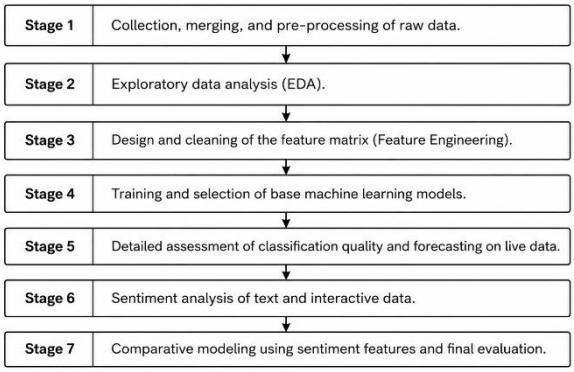


Fig. 6. **Conceptual model for the intelligent selection of crowdfunding project**
***Source*: compiled by the authors**

The Decision Support Workflow is examined from the perspective of the decision-maker (DM) regarding the selection and funding of crowdfunding projects, based on the following sequence of steps (Fig. 7):

Stage 1: The DM enters data on new active (live) projects into the system for an initial automated screening of their basic characteristics.

Stage 2: The system analyses the project parameters (funding target, duration, category) and alerts the DMA to any anomalies or critical constraints identified at the outset.

Stage 3: The DM receives an automatically generated assessment of the project's probability of success based on the best algorithm and classifies candidates according to the model's reliability score.

Stage 4: The decision-maker examines the results of the VADER and FinBERT linguistic analyses to identify 'excessive emotional optimism' on the part of authors in the short description (blurb), which often indicates an attempt to conceal poor quality and is a sign of the risk of failure.

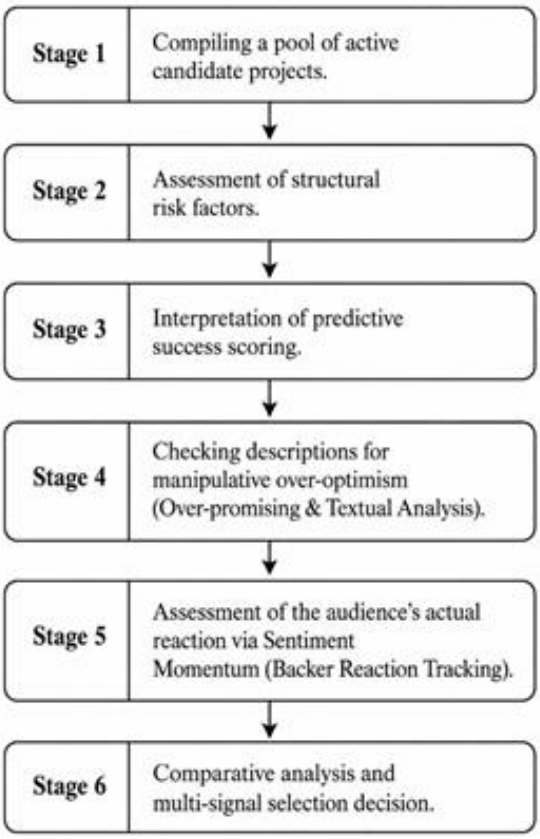


Fig. 7. **Decision support workflow**
***Source*: developed by the authors**

Stage 5: The Decision-Making Body tracks changes in the tone of backers' comments (whether positive sentiment is rising or falling), using the community's reaction as an independent tool to confirm the project's quality.

Stage 6: By comparing the model's predictive assessments with backers' behavioral signals over time, the decision-maker makes a final, informed decision on whether to support or reject the project.

The implementation of intelligent systems in the financial sector is often hampered by the 'black box' problem. Modern, highly efficient machine learning algorithms (such as random forests, gradient boosting, and multilayer perceptrons) can identify complex nonlinear patterns in data. However, the logic behind their decision-making remains opaque to the end user [12]. For investors and crowdfunding platforms, it is critically important to understand why the system recommends a particular project, as financial accountability and trust in the decision support worklow as a whole depend on this [14].

To address this issue, the concept of Explainable AI (XAI) has been integrated into the architecture of the proposed DSW. The use of interpretation methods enables complex mathematical relationships in ML models to be transformed into human-readable explanations [19].

The developed DSW architecture provides a comprehensive approach to selecting crowdfunding projects by integrating machine learning, natural language processing, and explainable artificial intelligence. The system automates data collection and processing, accounting for both structured project characteristics and emotional signals derived from text descriptions and social activity, thereby improving the accuracy of funding success predictions. The use of machine learning algorithms ensures an accurate assessment of project prospects, whilst the application of SHAP and LIME methods enhances the transparency and interpretability of the results.

Experimental testing and practical validation

For experimental testing and practical validation of the developed intelligent DSW, anonymized data from the Kickstarter crowdfunding platform were selected as the empirical basis [34]. Several key factors justify the choice of the Kickstarter Dataset. Firstly, Kickstarter is the global leader in reward-based crowdfunding, ensuring a highly representative sample across various technological, creative, and social sectors. Secondly, this dataset combines both rigidly structured metrics (financial targets, timeframes, and geographical location) and rich unstructured text data (project titles, short descriptions–blurbs), opening up opportunities for the application of advanced natural language processing (NLP) and sentiment analysis methods [2].

The application of machine learning (ML) and deep learning (DL) methods for predictive analysis enables the transformation of loosely structured datasets into actionable strategic insights, thereby reducing financial risks and optimizing the allocation of limited resources.

Results of the baseline models

Training on 40 structured features without sentiment yielded the following results on the test sample (n=58,316, Table 2). Confusion Matrix for RF: TN=15,805, FP=7,938, FN=4,416, TP=30,157; for MLP: TN=16054, FP=7689, FN=5062, TP=29511; for Decision Tree: TN=13327, FP=10416, FN=3937, TP=30636; for KNN: TN=14730, FP=9013, FN=5216, TP=29357; for Logistic Regression: TN=15064, FP=8679, FN=5538, TP=29035.

Random Forest achieved the best performance: F1=83.00 %, AUC=86.34 %, Balanced Accuracy=76.90%. MLP came second (F1=82.23 %, AUC=85.47 %). The three models (LR, KNN, and DT) clustered around F1≈80-81%, confirming the non-linear nature of the task. The difference between RF and LR was 2.67 % for F1 and 4.11 % for AUC.

The ranking of all five ML models is shown in Fig. 8 and Fig. 9 using F1-score and ROC correspondingly.

The impact of Sentiment Analysis on classification quality

Adding VADER and FinBERT features did not yield a statistically significant improvement in classification metrics (Table 3). The greatest degradation is observed for KNN (ΔF1 = -0.75% with FinBERT), which is explained by the curse of dimensionality when adding irrelevant features. RF demonstrates the highest robustness (ΔF1 = -0.03%) due to its ensemble nature.

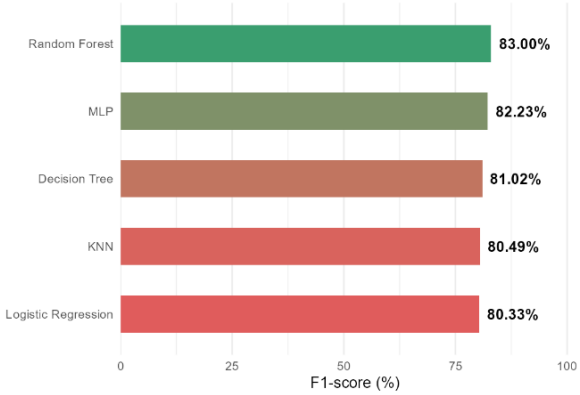


Fig. 8. **F1-score ranking of ML models**
***Source*: developed by the authors**

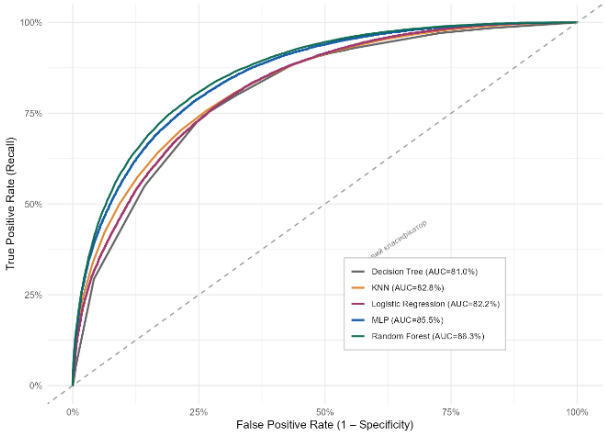


Fig. 9. **ROC curves ranking of ML models**
***Source*: developed by the authors**

Table 2. **Metrics for baseline models (n=58,316, Baseline: 40 features)**

Model	Accuracy, %	Precision, %	Recall, %	F1, %	AUC, %	Spec., %	Bal.Acc., %
Random Forest	**78.82**	79.16	87.23	**83.00**	**86.34**	**66.57**	**76.90**
MLP	78.13	79.33	**85.36**	82.23	85.47	67.62	76.49
Decision Tree	75.39	**74.63**	88.61	81.02	80.97	56.13	72.37
KNN (k=21)	75.60	76.51	84.91	80.49	82.76	62.04	73.48
Logistic Reg.	75.62	76.99	83.98	80.33	82.23	63.45	73.71

***Source*: calculated by the authors**

Table 3. **Comparison of F1-scores across three feature configurations**

Model	Baseline F1, %	+VADER F1, %	+FinBERT F1, %	ΔF1 (FinBERT), %	ΔAUC, %
Random Forest	83.00	82.92	82.97	-0.03	+0.11
MLP	82.23	81.75	81.82	-0.41	-0.35
Decision Tree	81.02	81.02	81.03	+0.01	+0.04
KNN	80.49	79.93	79.74	-0.75	-1.10
Logistic Reg.	80.33	80.30	80.29	-0.04	+0.01

***Source*: calculated by the authors**

It is important that FinBERT features rank 7-9 in the random forest (RF) importance ranking (Gini Importance: finbert_negative=3,130, finbert_neutral=3,067, finbert_positive=3,051), trailing structural predictors. Top 6 features: prep_days (5,885), goal (5,038), log_goal (4,875), staff_pick (4,384), duration_days (4,266), launch_year (3,727) (Fig. 10).

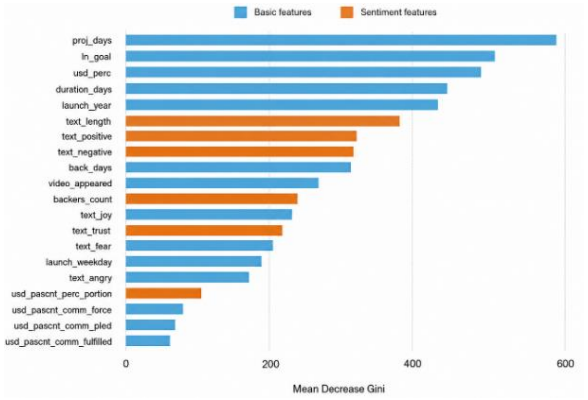


Fig. 10. **Top-20 features by mean decrease in Gini importance for Random Forest**
***Source*: developed by the authors**

Thus, FinBERT is indeed used by the model, but stronger structural signals overshadow its contribution.

Because ‘staff_pick’ (the platform’s editorial ‘Projects We Love’ label) is a strong predictor (Gini ≈ 4,384, rank 4) and is an editorial signal that may be assigned at or during a campaign rather than being guaranteed at the launch instant, its influence was assessed by an ablation. Re-training the Random Forest on the baseline features under the identical 70/30 split, seed, and settings, but excluding ‘staff_pick’, reproduced the published full-feature result (F1 = 83.0 %, AUC = 86.3 %) and yielded F1 = 81.52 % and AUC = 83.61 % without the label – a modest decline of ΔF1 = -1.51 % and ΔAUC = -2.79 %. The model therefore relies on ‘staff_pick’ only partially: it remains a strong classifier without it (F1 > 81 %, AUC > 83 %) and degrades gracefully should the label be unavailable at prediction time.

The phenomenon of over-promising and sentiment analysis by class

An analysis of the mean values of sentiment features across target classes revealed a counterintuitive result (Table 4): failed projects (target=0) are characterized by higher average VADER compound scores (0.288 vs 0.256) and FinBERT probabilities (0.131 vs 0.119) than successful ones (target=1).

This phenomenon has been termed ‘over-promising’: failed projects show, on average, slightly higher VADER compound (0.288 vs 0.256) and FinBERT positive (0.131 vs 0.119) scores than successful ones. This difference is highly significant given the large sample ($p < 0.001$). However, the differences were of negligible magnitude (Cohen’s d ≈ -0.08 and -0.10, respectively), so it is a weak association rather than evidence that optimistic wording causes failure or that backers act on it. The association is consistent with signal theory [7] and offers a plausible reason why adding sentiment features does not improve, and occasionally even slightly worsens, classification performance.

Consistency check: blurb sentiment vs. backers’ comment sentiment

As an auxiliary consistency check – not a validation of the sentiment models’ accuracy, since the same automatic tools produce both signals – a correlation analysis was carried out between the tone of authors’ descriptions (prior to launch) and the tone of backers’ comments (after launch) across the 2,085 overlapping projects (1,637,732 comments on those projects – 29.4 % of the 5,571,285-comment corpus). The results are presented in Table 5.

A statistically significant but weak correlation (Fig. 11) was found between blurb sentiment and comments (Pearson's r = 0.069, p = 0.0017). The average VADER score for comments (0.337) is significantly higher than that for blurbs (0.265), reflecting the natural positive bias inherent in live communication.

Table 4. **Average values of sentiment features by class (data set, n=194,388; target=0 (failure, n=78234); target=1 (success, n=116,154))**

Class	VADER compound	VADER pos	VADER neg	FinBERT pos	FinBERT neg
target=0	0.288	0.142	0.035	0.131	0.035
target=1	0.256	0.131	0.041	0.119	0.041

***Source*: calculated by the authors**

Table 5. **Correlation analysis: Blurb sentiment vs. comment sentiment**

Analysis	Coefficient	p-value	95 % CI	Significance
Pearson r (blurb vs comments VADER)	r = 0.069	0.0017	[0.026; 0.111]	Yes (α=0.05)
Spearman ρ (blurb vs comments VADER)	ρ = 0.062	0.0045	-	Yes (α=0.05)
Pearson r (blurb VADER vs number of comments)	r = 0.032	0.1462	-	No
Pearson r (blurb vs sentiment momentum)	r = 0.054	0.0140	-	Yes (α=0.05)

Source: **calculated by the authors**

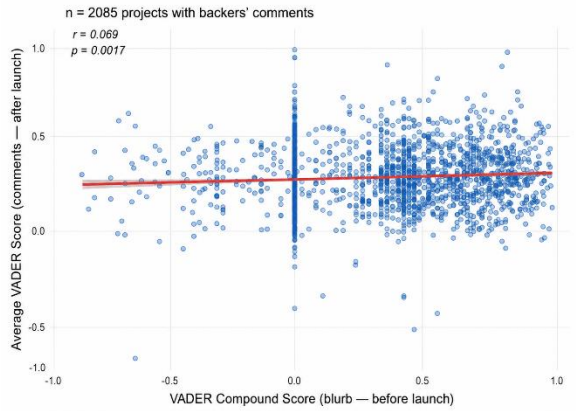


Fig. 11. **Correlation between blurb sentiment and comments**
***Source*: developed by the authors**

Sentiment momentum (the difference between later and earlier comments) also correlates significantly with blurb sentiment (r = 0.054, p = 0.014): a more positive authorial text is associated with a consistently positive audience reaction throughout the campaign. The weak correlation (r≈0.07) is explained by the fact that backers' reactions are primarily shaped by the product's actual characteristics rather than the marketing text.

Assessment and forecasting of Live projects

Random Forest (Baseline) was applied to 2,042 live projects (Fig. 12).

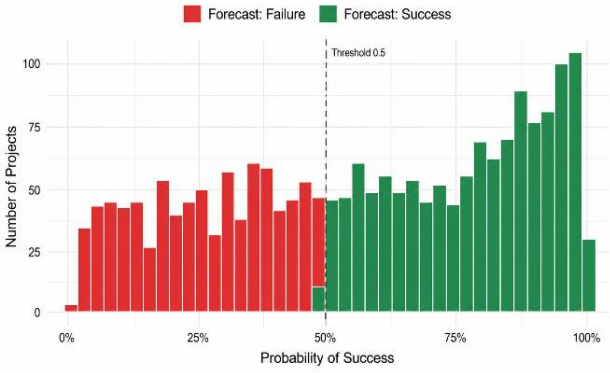


Fig. 12. **Forecasting of live projects success**
***Source*: developed by the authors**

Distribution of predictions: 'High confidence' (P > 0.80): 633 projects (31.0%); 'Medium' (0.60-0.80): 391 (19.1 %); 'Low'(0.40-0.60): 369 (18.1%); 'Very low' (P<0.40): 649 (31.8%). In total, 1,222 out of 2,042 projects (59.8 %) were predicted to be 'successful' – this matches the overall class distribution in the dataset (59.8 %), which reflects agreement with the base rate rather than a formal calibration check – the latter would require comparing predicted probabilities with realized outcomes, which are not yet available for live projects. Among the top performers in the prediction: 'MONSTER MATADOR: BLOOD OF THE MATADOR #1-2'(P=1.000), 'Set a Watch: Winterlands' (P=1.000). The underperformers are: 'Mini & Me Play Cafe' (P=0.000) and 'VidTube-The Video Platform' (P=0.013).

DISCUSSION

The results obtained allow us to answer three research questions.

RQ1 (Does sentiment analysis of blurbs improve crowdfunding success prediction?): No, in a statistically significant sense. The VADER and FinBERT sentiment features did not improve the F1 score of any model by more than 0.1 %. A consistent interpretation is the weak 'over-promising' association: blurb sentiment is, if anything, a slight inverse correlate of success rather than a direct one, so it adds little predictive value.

RQ2 (Optimal algorithm for DSW): Random Forest is the undisputed winner (F1=83.00 %, AUC=86.34 %) thanks to its combination of the highest accuracy, robustness to irrelevant features, and interpretability via Gini importance. MLP comes in a close second (F1=82.23 %), but requires greater computational resources and is less interpretable. LR is recommended as a fast linear baseline (F1=80.33 %, training <3 seconds).

RQ3 (Correlation between blurbs and comments?): Weak but significant (r=0.069, p=0.0017). This result indicates that blurb sentiment is weakly associated with the audience's later reaction, but is not the primary predictor of financial success. The comment corpus, therefore, serves as an auxiliary consistency signal rather than as a means of validating the sentiment models.

The most important structural predictor turned out to be prep_days (Gini = 5,885) – the time taken

to prepare a campaign, from creating a draft to launch. This result is consistent with signal theory [7]: a lengthy preparation period is a 'costly signal', indicating the seriousness of the author's intentions. The second most important predictors are goal and log_goal (Gini≈5,038/4,875), which confirm the classic findings [10] regarding the critical role of realistic financial expectations.

Limitations of the study: (1) comments are available only for 2,085 (≈1%) projects, all of which are successful; (2) FinBERT did not have access to update texts or full page descriptions; (3) SHAP and LIME analyses were not introduced as additional methods for feature importance. Future directions: integration of full campaign page texts, analysis of sentiments' comment time series for the dynamic phase of the DSW, and application of Gini importance for detailed interpretation of RF decisions.

Let us consider the integrated indicator 'Iattr' – the project investment attractiveness index Iattr = 0.50·PML + 0.30·Ssent + 0.20·Qqual–to ensure flexible adaptation of the DSW to the investment strategy.

This formula combines mathematical model forecasts with expert assessments and community feedback:

1. Iattr (Investment Attractiveness Index)–the overall investment attractiveness index of a start-up or project. Based on its value, the investor decides whether or not to fund the project.

2. PML (Probabilistic Machine Learning) – the probability of the project's success, as predicted by the best machine learning algorithm (in the study, this is the Random Forest model with AUC of over 86%), based on initial data (goal, category, country) and an analysis of the text descriptions.

3. Ssent (Sentiment Score)–a composite sentiment indicator reflecting the reaction of the audience and backers in the comments (based on the analysis of metrics such as VADER compound or FinBERT). If the community is optimistic and supports the project in the comments, this indicator boosts the overall index.

4. Qqual (Qualitative/Expert Score)–a qualitative assessment of the project, assigned manually by experts or investors based on specific criteria that are difficult to capture using algorithms (for example, the team's reputation, technological innovation, the feasibility of the prototype, etc.).

Weighting factors are set as benchmark at the following levels to balance the system: 50% depends on the mathematical PML forecast; 30% is based on emotional sentiment and social proof from the first potential investors; 20% of success is determined by expert analysis. This combination will enable the system to minimize the risks of information asymmetry, where authors attempt to embellish their project (the phenomenon of 'over-promising').

To make the index reproducible, all three components are first rescaled to a common [0, 1] range before weighting: PML is already a probability in [0, 1]; Ssent is obtained by min–max normalization of the chosen sentiment metric to [0, 1] (for VADER, Ssent = (compound + 1)/2; for FinBERT, the positive-class probability is used directly); and Qqual is an expert rating mapped to [0, 1]. The weights 0.50 / 0.30 / 0.20 are expert-assigned design parameters rather than empirically optimized coefficients: they deliberately give the largest weight to the data-validated ML forecast (PML), a smaller weight to the softer community-sentiment signal (Ssent), and the least to the subjective expert score (Qqual); they are tunable to the investor's strategy and are offered as a proposed heuristic, not as a result derived from the present data. Because Ssent (backers' comments) and Qqual (expert input) are not available at the initial, pre-launch selection stage, the index reduces to PML; Ssent and Qqual are intended to refine the ranking once the campaign is live and expert input has been collected.

CONCLUSIONS

The study achieved its objective – the development and experimental validation for the intelligent selection of crowdfunding projects on a large-scale dataset (194,388 Kickstarter projects from 2009 to 2026). Key results:

1. A Random Forest model based on basic 40 structured features achieved an F1 score of 83.00% and an AUC of 86.34 % – the best result among the five algorithms studied. The prediction for 2,042 live projects (59.8% 'successful') matches the overall class distribution; this base-rate agreement is not a formal calibration assessment, since the live outcomes are not yet known.

2. The VADER and FinBERT sentiment features do not significantly improve the binary 'success/fail' classification (ΔF1≈±0-0.75%). FinBERT ranks 7th–9th in terms of importance (Gini≈3,050-3,130), trailing behind structural predictors (prep_days, goal, staff_pick).

3. The phenomenon of 'over-promising' has been identified: failed projects have a higher VADER compound score (0.288 vs 0.256) and a higher FinBERT positive score (0.131 vs 0.119). The difference is statistically significant but of

negligible effect size (Cohen's d ≈ −0.08…−0.10), i.e., a weak association rather than a causal effect; higher linguistic optimism is, at most, a weak negative signal of quality.

4. A consistency check across 1,637,732 backer comments on the 2,085 overlapping projects revealed a weak but statistically significant correlation between blurb sentiment and comment sentiment (Pearson r = 0.069, p = 0.0017); this is an association between two automatic signals and does not, by itself, validate the accuracy of the sentiment models.

5. For practical implementation in a decision support workflow, Random Forest is recommended as the optimal compromise between accuracy, stability, and interpretability via Gini importance. The composite metric enables flexible adaptation to the investment strategy when deciding on project funding.

Conflict of interests: Authors declare that they do not have conflict of interests, in particular financial, personal, copyright or any of a different nature, which could have influenced the research, as well as the results published in this articles

Financing: Research was carried out without financial support

Accessibility data: All data available in digital or graphic form in the main text manuscript. The processed dataset supporting this study is openly available in Zenodo at https://doi.org/10.5281/zenodo.22131782

Using artificial intelligence tools: During preparation manuscript used tool Grammarly for grammatical and stylistic text checks. All fragments processed by him were checked and edited by the author. AI was not used to generate text, formulas, and tables, formation scientific provisions, receipt results or formulation conclusions. Scientific the provisions, results and conclusions are our own, intellectual by the contribution of the authors

REFERENCES

1. Gazzawe, F. "A system-oriented machine learning approach for planning and execution of decisions in software project management". *International Journal of Advanced Computer Science and Applications*. 2026; 17 (2): 649–659. DOI: https://doi.org/10.14569/ijacsa.2026.0170265.

2. Avvad, H. "Crowdfunding success prediction using machine learning: A comparative study based on Türkiye's campaigns". *Black Sea Journal of Engineering and Science*. 2026; 9 (2): 646–663. DOI: https://doi.org/10.34248/bsengineering.1772673.

3. Kobets, V. & Berehovyi, O. "The development of investment projects attractiveness assessment software". *Communications in Computer and Information Science*. 2022; 1635: 52–66. DOI: https://doi.org/10.1007/978-3-031-14841-5_4.

4. Kilinich, D. & Kobets, V. "Support of investors' decision making in economic experiments using software tools". *CEUR-WS*. 2019. p. 277–288. DOI: https://doi.org/10.55056/ceur-ws.org/vol-2393/paper_273.pdf.

5. Kobets, V., Petrov, O. & Koval, S. "Sustainable robo-advisor bot and investment advice-taking behavior". *Lecture Notes in Business Information Processing*. 2022; 465: 15–35. DOI: https://doi.org/10.1007/978-3-031-23012-7_2.

6. Kobets, V. "Evaluation of investment portfolio by application of multi-criteria decision making methods using robo-advisor". *Proceedings on Engineering Sciences*. 2022; 4 (3): 301–312. DOI: https://doi.org/10.24874/pes04.03.008.

7. Li, L., Yang, L., Zhao, M., Liao, M. & Cao, Y. "Exploring the success determinants of crowdfunding for cultural and creative projects: An empirical study based on signal theory". *Technology in Society*. 2022; 70: 102036. DOI: https://doi.org/10.1016/j.techsoc.2022.102036.

8. Alam, M. "Multi-theoretic perspective of Backer decision-making in crowdfunding: Interplay of Backer, Founder and AI". *Doctoral dissertation, The University of Texas Rio Grande Valley*. 2025. – Available from: https://scholarworks.utrgv.edu/etd/1748. – [Accessed: 20.04.2026].

9. Qiu, Z., Qu, Y., Yang, S., Zhang, W., Xu, W. & Zhao, H. "Enhancing startup financing success prediction based on social media sentiment." *Systems.* 2025; 13 (7): 520, D OI: https://doi.org/10.3390/systems13070520.

10.Costantino, F., Di Gravio, G. & Nonino, F. "Project selection in project portfolio management: An artificial neural network model based on critical success factors". *International Journal of Project Management*. 2015; 33 (8): 1744–1754. DOI: https://doi.org/10.1016/j.ijproman.2015.07.003.

11.Mitrovic, Z. M., Rakicevic, A. M., Petrovic, D. C., Mihic, M. M., Rakicevic, J. D. & Jelisic, E. T. "Systems thinking in software projects-an artificial neural network approach". *IEEE Access*. 2020; 8: 213619–213635. DOI: https://doi.org/10.1109/access.2020.3040169.

12.Feng, Y. & Park, J. “Identifying and predicting key industry–education integration projects via explainable artificial intelligence (XAI) based machine learning models”. *PeerJ Computer Science*. 2026; 12: e3868. DOI: https://doi.org/10.7717/peerj-cs.3868.

13.Oh, J. & Ashuri, B. “Explainable AI-driven expert system for impact profiling of contract modifications in construction projects”. *Expert Systems with Applications*. 2026; 318: 132030. DOI: https://doi.org/10.1016/j.eswa.2026.132030.

14. Amanzholova, S., Mutanov, G., Ussatova, O., Aldasheva, L., Amirova, A. & Naumenko, V. “A conceptual framework for AI- and Blockchain-Enabled research project evaluation systems”. *Information*. 2026; 17 (2): 151. DOI: https://doi.org/10.3390/info17020151.

15.Ma, J., Zheng, Z., Zhu, P. & Liu, Z. “A machine learning and large language model-integrated approach to research project evaluation”. *Journal of Database Management*. 2024; 35 (1): 1–14. DOI: https://doi.org/10.4018/jdm.345400.

16.Uddin, S., Yan, S. & Lu, H. “Machine learning and deep learning in project analytics: methods, applications and research trends”. *Production Planning & Control*. 2024; 36 (7): 873–892. DOI: https://doi.org/10.1080/09537287.2024.2320790.

17.Jiang, C., Pérez-Estébanez, R. & Urquía-Grande, E. “Is equity crowdfunding the leapfrog to companies’ success? Financial performance in China”. *Computational Intelligence and Neuroscience*. 2022; 7814550: 1–12. DOI: https://doi.org/10.1155/2022/7814550.

18.Elitzur, R. Katz, N. Muttath, P. & Soberman, “The power of machine learning methods to predict crowdfunding success: Accounting for complex relationships efficiently”. *Journal of Business Venturing Design*. 2024; 3: 100022. DOI: https://doi.org/10.1016/j.jbvd.2024.100022.

19.Waqar, A. “Intelligent decision support systems in construction engineering: An artificial intelligence and machine learning approaches”. *Expert Systems with Applications*. 2024; 249: 123503. DOI: https://doi.org/10.1016/j.eswa.2024.123503.

20.Liu, Y., Zhang, K., Xue, W. & Zhou, Z. “Crowdfunding innovative but risky new ventures: the importance of less ambiguous tone”. *Financial Innovation*. 2024; 10 (1). DOI: https://doi.org/10.1186/s40854-023-00529-8.

21.You, H. W., Usman, S. M., Yap, B. W., Muhammad, H. & Bukhari, F. a. S. “Prediction of crowdfunding success in Malaysia using quality signaling theory and machine learning”. *Business Process Management Journal*. 2025. p. 1–30. DOI: https://doi.org/10.1108/bpmj-02-2025-0175.

22.Chaudhary, S. (2024). “Enhancing success prediction in crowdfunding platform using LIME-explainable AI method”. *2024 IEEE Students Conference on Engineering and Systems: Interdisciplinary Technologies for Sustainable Future, SCES*. 2024. p. 1–6. DOI: https://doi.org/10.1109/sces61914.2024.10652343.

23.Tang, H., Zhao, Y., Li, N. & Zhu, J. “Enhancing crowdfunding financing speed through image-text integration strategies: Insights from information processing theory and explainable machine learning methodology”. *Technology in Society*. 2025; 84: 103136. DOI: https://doi.org/10.1016/j.techsoc.2025.103136.

24.Anglin, A. H., Short, J. C., Drover, W., Stevenson, R. M., McKenny, A. F. & Allison, T. H. “The power of positivity? The influence of positive psychological capital language on crowdfunding performance”. *Journal of Business Venturing*. 2018; 33 (4): 470–492. DOI: https://doi.org/10.1016/j.jbusvent.2018.03.003.

25.Biao, Y., Yu, Y., Chuang, H. & Chang, Y. “An analysis on the importance of persuasion strategies in Environmental-Oriented online crowdfunding projects”. *Lecture Notes in Networks and Systems*. 2023; 739: 1130–1142. DOI: https://doi.org/10.1007/978-3-031-37963-5_78.

26.Chan, C. R., Pethe, C. & Skiena, S. “Natural language processing versus rule-based text analysis: Comparing BERT score and readability indices to predict crowdfunding outcomes”. *Journal of Business Venturing Insights*. 2021; 16: e00276. DOI: https://doi.org/10.1016/j.jbvi.2021.e00276.

27.Shen, Y. & Zhang, P. K. “Financial sentiment analysis on news and reports using large language models and FinBERT”. In *IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*. 2024. p. 717–721. DOI: https://doi.org/10.1109/icpics62053.2024.10796670.

28.Peng, B., Chersoni, E., Hsu, Y. & Huang, C. “Is domain adaptation worth your investment? Comparing BERT and FinBERT on financial tasks”. In *Proceedings of the 3rd Workshop on Economics and Natural Language Processing, ECONLP*. 2021. p. 37–44. DOI: https://doi.org/10.18653/v1/2021.econlp-1.5.

29.Yang, J., Xin, J., Zeng, Y. & Liu, P. J. “Signaling and perceiving on equity crowdfunding decisions – a machine learning approach”. *Small Business Economics*, 2025; 65 (1): 315–356. DOI: https://doi.org/10.1007/s11187-024-00991-3.

30.Wang, W., Zhu, K., Wang, H. & Wu, Y. J. “The impact of sentiment orientations on successful crowdfunding campaigns through text analytics”. *IET Software*. 2017; 11 (5): 229–238. DOI: https://doi.org/10.1049/iet-sen.2016.0295.

31.Geng, S., Huang, M. & Wang, Z. “A swarm enhanced light gradient boosting machine for crowdfunding project outcome prediction”. *Lecture Notes in Computer Science*. 2020. p. 372–382. DOI: https://doi.org/10.1007/978-3-030-62463-7_34.

32.Hutto, C. J. & Gilbert, E. E. “VADER: A parsimonious rule-based model for sentiment analysis of social media text”. *Proceedings of the International AAAI Conference on Web and Social Media*. 2014; 8 (1): 216–225. DOI: https://doi.org/10.1609/icwsm.v8i1.14550.

33.Araci, D. “FinBERT: Financial sentiment analysis with pre-trained language models”. *arXiv*. 2019. DOI: https://doi.org/10.48550/arXiv.1908.10063.

34.Starchenko, G. & Kobets, V. “Kickstarter crowdfunding dataset for machine-learning and sentiment-based prediction of project success (2009–2026) [Data set]”. *Zenodo*. 2026. DOI: https://doi.org/10.5281/zenodo.22131782.

Received 15.09.2026
Received after revision 23.10.2026
Accepted 30.10.2026

DOI: https://doi.org/10.15276/hait.09.2026.40
УДК 004.8:65.012.12

Машинне навчання та аналіз настроїв для інтелектуального відбору проєктів: розробка та експериментальна валідація

Кобець Віталій Миколайович[1)]
ORCID: https://orcid.org/0000-0002-4386-4103; vkobets@kse.org.ua. Scopus Author ID: 56006224700
Старченко Григорій Володимирович[1), 2)]
ORCID: https://orcid.org/0000-0003-2707-1055; g.v.starchenko@gmail.com. Scopus Author ID: 57220892211
Арданов Павло Євгенович[3), 4)]
ORCID: https://orcid.org/0000-0003-0229-915X; pavlo.ardanov@gmail.com. Scopus Author ID: 54381559600
[1)] Херсонський державний університет, вул. Університетська, 27. Херсон, 73003, Україна
[2)] Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського” Київ, Берестейський проспект, 37, 03056, Україна
[3)] Інститут біомедичних технологій, Відкритий міжнародний університет розвитку людини «Україна», вул. Львівська, 23 Київ, 03115, Україна
[4)] Університет Оулу, Вулиця Пентті Кайтера, 1, Оулу, 90570, Фінляндія

АНОТАЦІЯ

Актуальність. Стрімке зростання кількості краудфандингових кампаній на таких платформах, як Kickstarter, у поєднанні з інформаційною асиметрією між засновниками та інвесторами, зумовлює потребу в інтелектуальних інструментах підтримки прийняття рішень, які враховують настрої інвесторів, а не покладаються виключно на структуровані фінансові показники. **Мета**. Розробити та експериментально перевірити робочий процес підтримки прийняття рішень для прогнозування успіху краудфандингових кампаній, який поєднує кілька алгоритмів класифікації з аналізом настроїв, використовуючи коментарі спонсорів як допоміжну перевірку узгодженості. **Завдання**. Формалізувати гібридний набір ознак, що поєднує структуровані метадані проєкту з текстовими індикаторами настроїв; навчити та порівняти алгоритми класифікації для різних конфігурацій ознак; оцінити вплив настроїв на якість прогнозування; інтерпретувати відмінності в настроях між успішними та невдалими кампаніями; звірити узгодженість настроїв анотацій із настроями в коментарях; а також прогнозувати результати для поточних кампаній. **Методи**. Дані з Kickstarter були оброблені шляхом видалення повторів, видалення винятків, логарифмічного перетворення та кодування «one-hot». П’ять алгоритмів класифікації – логістична регресія, k-найближчих сусідів, дерево рішень, «випадковий ліс» та багатошаровий перцептрон – були навчені у трьох конфігураціях, що поєднували

структуровані ознаки з лексикографічними та трансформерними оцінками настроїв; кореляції Пірсона та Спірмена пов'язували настрої анотацій та коментарів. **Результати**. Було проаналізовано близько 194000 очищених проєктів (2009-2026 рр.) та 1,64 млн проскорованих коментарів на 2 085 спільних проєктах. Найкращі результати показав «випадковий ліс»; ознаки сентименту не дали суттєвого поліпшення. Кампанії, що зазнали невдачі, демонстрували систематично вищі показники сентименту, ніж успішні – це ефект «надмірних обіцянок» (слабка асоціація, а не причинний зв'язок). Перевірка узгодженості між проєктами виявила слабку, але статистично значущу кореляцію між сентиментом анотацій та коментарів. **Наукова новизна**. Настрій формалізується не як прямий предиктор, а як зворотний сигнал якості в рамках гібридного, пояснюваного робочого процесу підтримки прийняття рішень, який спільно моделює структуровані метадані, текстовий настрій та відгуки спільноти, використовуючи коментарі спонсорів як допоміжний сигнал узгодженості. **Практичне значення**. Цей робочий процес може допомогти краудфандинговим платформам та інвесторам завдяки автоматизованому відбору кампаній, інтерпретації рекомендацій на основі характеристик та комплексному індексу інвестиційної привабливості, що поєднує прогнози моделі, сентимент та експертну оцінку. **Висновки**. Цей робочий процес забезпечує точний і зрозумілий відбір краудфандингових проєктів. Водночас компонент аналізу настроїв потребує подальшої валідації з використанням повних текстів кампаній та часових рядів коментарів, а також тестування на інших платформах.

Ключові слова: інтелектуальний відбір проєктів; машинне навчання; аналіз настроїв; краудфандинг; робочий процес підтримки прийняття рішень

ABOUT THE AUTHORS



Vitaliy M. Kobets - Doctor of Economic Science, Professor, Department of Computer Science and Software Engineering. Kherson State University. 27, Universitetska St. Kherson, 73003, Ukraine
ORCID: https://orcid.org/0000-0002-4386-4103; vkobets@kse.org.ua. Scopus Author ID: 56006224700
***Research field*:** Data Science in Economics; Evolutionary Microeconomics; Robo-Advisor

Кобець Віталій Миколайович – доктор економічних наук, професор кафедри Комп'ютерних наук та програмної інженерії. Херсонський державний ун-т, вул. Університетська, 27. Херсон, 73003, Україна
***Галузь досліджень*:** наука про дані в економіці; еволюційна мікроекономіка; автоматизований фінансовий консультант



Grygoriy V. Starchenko - Doctor of Economic Sciences, Professor, Department of International Business and Logistics. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37 Beresteiskyi Avenue, Kyiv, 03056, Ukraine; student of the master's degree program on Information Systems and Technology. Kherson State University. 27, Universitetska St. Kherson, 73003, Ukraine
ORCID: https://orcid.org/0000-0003-2707-1055; g.v.starchenko@gmail.com. Scopus Author ID: 57220892211
***Research field*:** Project Management; Digital Transformation; Artificial Intelligence and Machine Learning

Старченко Григорій Володимирович - доктор економічних наук, професор, професор кафедри Міжнародного бізнесу та логістики. Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Берестейський проспект (Перемоги), 37, 03056, Україна; студент спеціальності F6 Інформаційні системи та технології магістерського рівня кафедри комп'ютерних наук та програмної інженерії. Херсонський державний ун-т, вул. Університетська, 27. Херсон, 73003, Україна
Галузь досліджень: управління проєктами; цифрова трансформація; штучний інтелект та машинне навчання



Pavlo Ye. Ardanov - PhD in Plant Physiology, Associate Professor Institute of Biomedical Technologies, Open International University of Human Development "Ukraine", 23, Lvivska Str., Kyiv, 03115; University of Oulu, Pentti Kaiteran katu 1, Oulu, 90570, Finland
ORCID: https://orcid.org/0000-0003-2707-1055; pavlo.ardanov@gmail.com. Scopus Author ID: 54381559600
***Research field*:** diversified cropping systems; plant-microbe interactions; agro-economic projects

Арданов Павло Євгенович - доктор філософії в галузі фізіології рослин, доцент Інституту біомедичних технологій Відкритого міжнародного університету розвитку людини "Україна", вул. Львівська, 23, Київ, 03115**;** Університет Оулу, вул. Пентті Кайтеран, 1, Оулу, 90570, Фінляндія
Галузь досліджень: диверсифіковані системи землеробства; взаємодія рослин і мікроорганізмів; агроекономічні проєкти