Received: 19 April 2026 | Revised: 8 June 2026 | Accepted: 23 July 2026 | Published online: 8 September 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–18
DOI: 10.47852/bonviewJDSIS620210062

BON VIEW PUBLISHING

Community-Based Innovation for Poverty Alleviation: A Text-Based Audit of Area-Based Research Funding in Thailand

Wichidtra Sudjarid[1] and Pita Jarupunphol[2,*]

[1] *Department of Environmental Science, Sakon Nakhon Rajabhat University, Thailand*

[2] *Department of Digital Technology, Phuket Rajabhat University, Thailand*

Abstract: This study evaluates a Thai area-based research funding portfolio using an explainable multi-label text classification framework. The analysis used 629 project records containing project titles, innovation descriptions, funding-source metadata, year, and binary poverty-related labels. Seven labels with sufficient support were retained for supervised modeling: Lack_Skills, Soil_Issues, Water_Shortage, Tech_Adaptation_Issues, Disaster_Risk, Vulnerable_Groups, and Welfare_Access. Because the corpus was small, the project texts were short, and the label distribution was highly imbalanced, the study used sparse Term Frequency-Inverse Document Frequency (TF-IDF) representations and interpretable classical machine-learning models rather than deep learning. The best held-out model was Linear Support Vector Machine (SVM), achieving macro-F1 = 0.5391 and weighted-F1 = 0.7798 under stratified five-fold cross-validation. Logistic Regression was second-best with macro-F1 = 0.5114 and weighted-F1 = 0.7582. The fold-level difference between Linear SVM and Logistic Regression was statistically significant (paired t-test p = 0.0021, Cohen's d = 3.1539; McNemar p = 0.0027), although the small number of folds requires cautious interpretation of effect-size magnitude. Ablation analysis showed that unigram-only text slightly outperformed the full feature set, while title-only modeling substantially reduced performance. Explainable lexical features were dominated by terms related to poverty, soil, water, disaster, and data. The findings support a practical role for explainable text analytics in portfolio screening, funding governance review, and evidence-informed monitoring of poverty-oriented research portfolios.

Keywords: explainable text classification, multi-label learning, research portfolio analysis, area-based development, poverty-oriented innovation

1. Introduction

Community-centered innovation, participatory research, place-based development, and inclusive innovation strategies are frequently used to justify geographically targeted investment to address local needs, build capabilities, and reduce poverty [1–5]. In that framing, the portfolio language of projects is not a trivial administrative field but a measurable record of how researchers define local problems and beneficiaries.

Thailand is often cited as a development success story, having moved from a low-income to an upper-middle-income country in less than a generation. However, this growth has not been spatially uniform. The "center-periphery" divide remains stark, with the Bangkok metropolitan area accounting for a disproportionate share of GDP. At the same time, the Northeastern region (Isan) continues to house most of the country's poor. This persistence of regional inequality, despite overall economic growth, is known as the "inequality paradox." Traditional top-down economic policies, infrastructure mega-projects, and industrial estates have often failed to trickle down to the grassroots level.

Development policy in Thailand and the broader Global South has shifted toward area-based development (ABD).

This approach argues that solutions to poverty cannot be centrally planned but must be contextualized to the specific cultural, geographical, and social fabrics of local communities. In parallel with this policy shift, the role of public research funding has evolved. Historically, research funds were allocated based on academic merit (scientific excellence). Today, there is an increasing demand for "Social Impact Research" projects designed specifically to solve societal problems.

This study addresses that need by framing funded projects as a multi-label text classification problem. Text classification surveys continue to distinguish between traditional sparse methods and deep learning approaches and also show that model choice depends on text length, label density, and the amount of available training data [6, 7]. Short-text studies in the social sciences further show that supervised machine-learning methods can remain competitive in small labeled corpora with infrequent categories, while requiring far less compute than deep models [8].

***Corresponding author:** Pita Jarupunphol, Department of Digital Technology, Phuket Rajabhat University, Thailand. Email: p.jarupunphol@pkru.ac.th

Rather than asking whether funding causes poverty reduction, the article asks whether the thematic structure of funded projects can be predicted from project language and a small set of administrative variables and which modeling choices produce the most defensible and interpretable results. That distinction matters: predictive performance speaks to portfolio pattern recognition, whereas causal claims would require treatment assignment, counterfactuals, and stronger control for confounders than this dataset provides.

The article is therefore positioned as an explainable portfolio audit. This framing is consistent with interpretability research that emphasizes audience-specific explanations and favors inherently interpretable models when the task is evidence-sensitive or policy-facing [9–11]. Recent surveys also stress that explanation methods should be evaluated for quality, audience fit, and decision context, especially in high-stakes settings [12–15].

1.1. Research objectives

The study has four objectives:

1) To assess whether poverty-related project categories can be predicted reliably from short project text and limited metadata.
2) To compare a defensible set of baseline, interpretable, and stronger classical machine-learning models.
3) To test whether preprocessing and metadata additions improve out-of-fold performance.
4) To interpret the most influential lexical features while keeping the discussion strictly grounded in observed results.

1.2. Research gap and contributions

Despite growing interest in community-based innovation, ABD, and poverty-oriented research funding, prior studies have rarely integrated portfolio-level policy analysis with reproducible multi-label text classification. Existing work commonly emphasizes conceptual policy discussion, descriptive portfolio review, or general machine-learning classification. Still, it does not fully combine short-text modeling, stratified cross-validation, ablation testing, paired statistical comparison, and explainable feature interpretation for auditing Thai area-based anti-poverty research funding. This study addresses that gap by treating project language as an auditable textual signal while maintaining a clear boundary between predictive portfolio screening and causal impact evaluation.

The study makes four contributions. First, it provides a policy-facing contribution by showing how funded research portfolios can be screened for poverty-related thematic structure using interpretable text analytics. Second, it provides a methodological contribution by evaluating multi-label classification under realistic low-resource conditions, where the corpus is small, the text fields are short, and several labels are sparse. Third, it contributes to explainability by identifying lexical cues that drive classification decisions and by interpreting these cues as portfolio-screening evidence rather than as causal mechanisms. Fourth, it contributes to research governance by showing how area-based funding agencies can use text-based auditing to monitor thematic concentration, neglected dimensions of poverty, and the alignment between funding mandates and project language.

The remainder of this article is organized as follows. Section 2 reviews literature on ABD, inclusive innovation, short-text classification, multi-label learning, and explainable machine learning. Section 3 describes the dataset, preprocessing workflow, pre-model diagnostics, feature engineering, target construction, model development, evaluation metrics, statistical validation, and ablation design. Section 4 presents the portfolio diagnostics, predictive model comparison, ablation results, statistical validation, and explainability findings. Section 5 discusses the theoretical, methodological, and policy implications of the findings, including limitations and future research directions. Section 6 concludes the study.

2. Literature Review

ABD posits that "place matters." Unlike sector-based policies (e.g., a national rice policy), ABD focuses on the nexus of problems within a specific geography.

1) The local trap: Scholars like Purcell [16] warn of the "local trap," the assumption that the local scale is inherently more democratic or effective. Without connection to national systems, local initiatives can become isolated and unsustainable.
2) Regional innovation systems (RIS): The ABD concept aligns with RIS theory, which emphasizes the interaction between local firms, universities, and government bodies. In Thailand, Program Management Unit-Area based (PMU-A) aims to act as the catalyst for these local systems, funding "Community Enterprises" or "One Tambol One Product (OTOP)" rather than individual firms.

Accordingly, community-based and inclusive innovation research suggests that innovation can address affordability, capability building, and poverty reduction, especially where development is organized around place-specific constraints [3, 5]. Place-based development research similarly emphasizes that local strategies are contextually distinct and should be evaluated with attention to local governance and uneven development, rather than relying solely on generic national averages [2]. The Thai area-based innovation-system literature echoes that point by identifying human capital, collaboration, capability, and sectoral linkages as central inputs [4].

Methodologically, short-text classification is a well-established task, but recent surveys still show that performance depends strongly on feature representation, label balance, and the computational budget available for training [6–8]. Logistic Regression remains a standard high-dimensional text classifier, and sparse term-weighting schemes remain a foundational representation for lexical models [17, 18]. Since this study is multi-label, the literature on binary relevance and related problem-transformation methods is especially relevant. Binary relevance is a representative decomposition strategy, and linear binary relevance can remain competitive while keeping the model family transparent and extensible [19, 20]. This technique can also be used to prioritize comparability, interpretability, and stability in the presence of sparse labels.

Evaluation in multi-label settings requires close attention, as accuracy can mask failures on minority labels. Reviews of classification metrics show that different metrics respond differently to label imbalance, and multi-label studies repeatedly recommend measures such as macro-F1 and Hamming loss as more informative than raw accuracy alone [21–24]. Explainability is therefore not decorative but methodological. Reviews of black-box explanations and interpretable machine learning emphasize that the meaning of an explanation depends on the audience and the decision context and that interpretable models are often preferable when the objective is accountability rather than post hoc storytelling [9–11].

Table 1
Comparative summary of prior studies

Study	Method	Data	Evaluation	Key gap
Israel et al. [1]	Review	Community-Based Participatory Research (CBPR) literature	Narrative synthesis	No text analytics
Hansasooksin and Tontisirin [4]	AHP	Thai expert survey	Criteria weighting	No predictive modeling
Mortazavi et al. [3]	Review	Bibliometric	Descriptive	No classification
Taha et al. [6]	Taxonomy + ML survey	Text	Empirical and experimental	Not policy-specific
Gasparetto et al. [7]	ML models	Mixed	Acc, F1	Limited policy use
Shyrokykh et al. [8]	ML models	Short text	F1	No reproducibility
Rudin [15]	XAI (conceptual)	Mixed	N/A	No benchmarking
This study	ML + ablation + stats	Short text	Macro-F1, CV	–

Note: Acc = accuracy; CV = cross-validation; ML = machine learning; XAI = explainable artificial intelligence.

As shown in Table 1, prior work either focuses on conceptual or descriptive analysis without predictive modeling or applies machine learning without rigorous validation and reproducibility. In contrast, the present study integrates explainable modeling, cross-validation, ablation analysis, and statistical testing within a unified framework, addressing a key methodological gap [14, 25].

3. Research Methodology

3.1. Data source and analytical framing

The dataset covered project records from 2020, 2021, and 2024. Records for 2022 and 2023 were not available in the analytical file. This temporal discontinuity is acknowledged as a limitation because funding priorities, project wording, and label distributions may vary across years. Accordingly, Year was retained only as contextual metadata and was not interpreted as evidence of a continuous temporal trend. The analysis should therefore be read as a cross-sectional and discontinuous portfolio audit rather than a longitudinal evaluation of policy change.

The dataset was sourced from the National Research Information and Innovation System (NRIIS) and related project-record documentation used for portfolio analysis. The unit of analysis was the funded project record, not an individual beneficiary, researcher, household, or community. Therefore, the study does not evaluate poverty reduction outcomes at the beneficiary level. Instead, it examines whether poverty-oriented thematic labels can be predicted from project language and limited administrative metadata.

The dataset contains 629 funded research records and 22 variables. The data combined structured metadata ("Funding_Source," "Year") with two text fields ("Research_Topic," "Technology_Innovation") and a multi-label set of binary socio-economic indicators. Because a single project could address multiple dimensions of poverty, the prediction task was formulated as multi-label text classification rather than as mutually exclusive single-label classification. Let

1) $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the dataset of funded projects,
2) $x_i \in \mathbb{R}^p$ denote the vectorized representation of the project i,
3) $y_i \in \{0,1\}^L$ denote a binary label vector over L poverty-related categories.

Because projects may address multiple socio-economic dimensions, each project is associated with a set of labels rather than a single class. The learning task is therefore defined as a multi-label classification problem, in which the model learns a mapping given by Equation (1).

$$f: \mathbb{R}^p \to \{0,1\}^L \tag{1}$$

3.1.1. Pre-model diagnostics

To characterize the portfolio structure before predictive modeling, the study conducted pre-model diagnostics that quantified (1) differences in label prevalence across funding agencies, (2) the strength of association between agency type and specific poverty indicators, and (3) co-occurrence patterns among multi-label categories. These diagnostics are descriptive and inferential with respect to the observed portfolio and are used to contextualize the subsequent text classification task rather than to support causal claims.

3.1.2. Prevalence estimation and confidence intervals

For each funding agency group g and label ℓ, prevalence was estimated as the observed proportion of projects with the label, where $x_{g\ell}$ is the number of projects in the group g positive for label ℓ and n_g is the number of projects in the group g, as in Equation (2).

$$\hat{p}_{g\ell} = \frac{x_{g\ell}}{n_g} \tag{2}$$

Uncertainty was summarized using a 95% confidence interval for a binomial proportion. Using the normal approximation for clarity (or an exact interval where required), the interval can be expressed as in Equation (3), with $z_{0.975} = 1.96$. These intervals support standardized comparison of agency-level portfolio emphases.

$$\hat{p}_{g\ell} \pm z_{0.975}\sqrt{\frac{\hat{p}_{g\ell}(1-\hat{p}_{g\ell})}{n_g}} \tag{3}$$

3.1.3. Association testing via Pearson's chi-square

To test whether label prevalence differs by funding agency type, contingency tables were constructed for each label ℓ. Let O_{ij} denote the observed count in row i (agency category) and column

j (label presence/absence), and let E_{ij} denote the expected count under independence, as in Equation (4).

$$E_{ij} = \frac{\left(\sum_j O_{ij}\right)\left(\sum_i O_{ij}\right)}{N} \tag{4}$$

Pearson's chi-square statistic is then utilized and explained in Equation (5) with degrees of freedom. $(r-1)(c-1)$, where r is the number of agency categories and $c = 2$ for presence/absence tables. This test evaluates whether observed differences in prevalence are larger than would be expected under random variation.

$$\chi^2 = \sum_i \sum_j \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{5}$$

3.1.4. Effect size via Cramer's V

Because large samples can yield small p-values for weak associations, effect sizes were reported using Cramer's V as in Equation (6), where N is the total sample size. This provides a scale-standardized measure of the strength of association between funding agency type and portfolio prioritization for a given label.

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(r-1,\, c-1)}} \tag{6}$$

3.1.5. Inter-label association (correlation) analysis

To examine co-occurrence structure among poverty indicators, pairwise associations between binary labels were summarized using Pearson correlation on the binary vectors (equivalent to the phi coefficient for binary pairs). For two labels a and b with binary realizations $\{y_{ia}\}$ and $\{y_{ib}\}$, the correlation is illustrated in Equation (7), where y_a and y_b are sample means.

$$r_{ab} = \frac{\sum_{i=1}^{N}(y_{ia} - \bar{y}_a)(y_{ib} - \bar{y}_b)}{\sqrt{\sum_{i=1}^{N}(y_{ia} - \bar{y}_a)^2}\sqrt{\sum_{i=1}^{N}(y_{ib} - \bar{y}_b)^2}} \tag{7}$$

The statistical significance of the correlation coefficients was evaluated using Equation (8).

$$t = r\sqrt{\frac{N-2}{1-r^2}},\ \text{with df} = N-2 \tag{8}$$

These pre-model diagnostics establish that the funding portfolio exhibits systematic structure rather than random variation, including strong agency-level specialization, pronounced label imbalance, and non-trivial co-occurrence among poverty dimensions. These properties motivate the subsequent modeling choices: a multi-label formulation, imbalance-aware evaluation, and the use of sparse, interpretable classifiers. The predictive models that follow therefore test whether this empirically observed structure, documented through descriptive and inferential statistics, can be recovered from project text and limited metadata, rather than implying any causal relationship between research funding and poverty alleviation.

3.1.6. Preprocessing pipeline

The processed text fields were short, with project language concentrated in titles and innovation descriptions. In addition to the previously reported mean token length, the revised analysis also reports the minimum, median, maximum, and mean token counts to describe the full distribution of text length. This addition clarifies the low-resource nature of the task and supports the decision to use sparse, interpretable lexical representations rather than data-hungry neural models. Because short administrative project text often contains compact but meaningful domain terms, TF–IDF was retained as a defensible baseline representation for small-sample text classification.

The original data were preserved unchanged. A derived analysis table was generated through safe file ingestion with encoding detection, duplicate verification, missing-value screening, label coercion to binary form, funding-source normalization, and text normalization. The text pipeline lowercased the input, removed punctuation, normalized whitespace, and concatenated the project title with the innovation description into a unified modeling field. Auxiliary variables, including token counts and per-record label counts, were generated to support diagnostics and visualization. The main modeling text was built by concatenating Research_Topic and Technology_Innovation.

3.1.7. Feature engineering

The primary representation was a TF–IDF feature space with unigram and bigram terms extracted from the cleaned text field. TF–IDF features were extracted from the normalized text because term-weighted lexical representations remain a strong baseline for text retrieval and categorization [6, 18]. The funding source was one-hot encoded, and the year was retained as a contextual covariate. Administrative context was incorporated through one-hot encoding of funding source and a scaled numeric representation of project year. This hybrid design was selected to balance interpretability, lexical coverage, and low-resource feasibility. Deep contextual embeddings were not introduced because the available environment did not support stable transformer training under the observed sample size and label imbalance. Let t be a term and d a document. The TF–IDF weighting is defined as in Equation (9), where $\text{tf}(t,d)$ is the term frequency of t in document d and N is the number of documents in the corpus.

$$\text{tfidf}(t,d) = \text{tf}(t,d) \cdot \log\left(\frac{N}{1 + |\{d' \in \mathscr{D} : t \in d'\}|}\right) \tag{9}$$

3.1.8. Target construction

Labels with fewer than 10 positive cases were retained for descriptive analysis but excluded from supervised modeling. This threshold was used to reduce unstable fold-wise estimation during five-fold cross-validation. With fewer than 10 positive cases, at least some validation folds would contain too few positive observations to reliably estimate label-wise performance. The project-level labels were harmonized into binary indicators using a positive-threshold rule (> 0 -> 1). One label anomaly, Tech_Adaptation_Issues = 2, was retained as a positive class during harmonization. Supervised modeling was restricted to labels with at least 10 positive cases: Lack_Skills, Soil_Issues, Water_Shortage, Tech_Adaptation_Issues, Disaster_Risk, Vulnerable Groups, and Welfare Access. Rarer labels were retained descriptively but excluded from supervised learning to avoid unstable fold-wise estimation. Welfare_Housing was constant zero. This decision improves the defensibility of supervised modeling, but it also limits the predictive analysis to labels with sufficient empirical support. Therefore, the results should not be interpreted as encompassing all possible dimensions of poverty in the funding portfolio.

3.1.9. Model development

The study used One-vs-Rest binary relevance because it is transparent, computationally stable, and appropriate for sparse multi-label settings with limited sample size. Each label-specific classifier can be inspected separately, which supports interpretability and policy-facing explanation. However, binary relevance assumes conditional independence among labels during model training and therefore does not directly model label co-occurrence. This is a known limitation in multi-label learning. In the present study, label dependencies were examined descriptively using inter-label correlation analysis, while the predictive models were deliberately kept simple and auditable. Future work may compare binary relevance with classifier chains, label-powerset methods, and neural multi-label architectures as larger, more balanced datasets become available.

Deep learning and pre-trained language models were not excluded because they are unimportant. They were excluded because the present corpus is small, project texts are short, and several labels have sparse support. Under these conditions, neural fine-tuning or prompt-based classification may be unstable without a larger labeled corpus, external validation, or careful domain adaptation. The purpose of this study is not to claim that TF–IDF is universally superior to pre-trained language models but to establish a transparent, reproducible, and proportionate baseline for low-resource portfolio auditing. Future studies with larger corpora should compare TF–IDF baselines against transformer embeddings, parameter-efficient fine-tuning, and few-shot prompting.

In this study, four supervised models were evaluated with a fixed random seed (42): One-vs-Rest Logistic Regression, One-vs-Rest Linear SVM, One-vs-Rest Random Forest, and XGBoost as the high-capacity fallback model. The multi-label problem was handled using independent binary classifiers for each label, consistent with binary relevance and the one-vs-all problem transformation [19, 20]. Logistic Regression and sparse linear classification are long-standing baselines for text categorization, whereas deep learning was omitted because the data are short-text, label-sparse, and limited in size [7, 8, 17]. In this case, two reference baselines were also evaluated: a majority-label baseline and a transparent keyword heuristic. Deep learning was not used because the corpus is small, the project texts are very short, and the label distribution is too sparse to justify neural fine-tuning [7, 8, 24]. For each label $\ell \in \{1, \ldots, L\}$, an independent binary classifier is trained as in Equation (10), where w_ℓ and b_ℓ are label-specific parameters.

$$\hat{y}_{i\ell} = \begin{cases} 1 & \text{if } w_\ell^\top x_i + b_\ell \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

The final prediction for the project i is the vector as described in Equation (11).

$$\hat{y}_i = (\hat{y}_{i1}, \ldots, \hat{y}_{iL}) \tag{11}$$

3.1.10. Evaluation and statistical validation

Because the task is multi-label and imbalanced, the revised evaluation reports Hamming loss and per-label F1 scores in addition to macro-F1, weighted-F1, and exact-match accuracy. Hamming loss indicates the average fraction of incorrect label decisions and is useful when each project can have multiple labels. Per-label F1 scores are reported to show whether performance is concentrated in frequent labels or distributed across the retained poverty categories. These additions improve the transparency of model performance and prevent the aggregate macro-F1 score from masking label-specific weaknesses.

Model performance was assessed using stratified five-fold cross-validation, with macro-averaged F1 score (macro-F1) as the primary evaluation metric and weighted-averaged F1 score (weighted-F1) as a secondary metric. Because standard label-wise stratification is not directly applicable in sparse multi-label settings, stratification was approximated by grouping samples by label cardinality (the number of active labels per project), which preserves a stable proxy for label-set complexity across folds. Additional diagnostics, including exact-match accuracy, Hamming loss, and per-label performance summaries, were monitored within the evaluation pipeline but are not emphasized in the main comparison.

Macro-F1 was prioritized because evaluation measures in multi-label classification respond differently to class imbalance, and accuracy-based metrics can be misleading in highly skewed label distributions [21–24]. This evaluation and validation strategy follows established best practices for classifier comparison, in which paired tests are preferred over independent-sample comparisons to control for shared data partitions and reduce variance due to fold assignment [26, 27]. For completeness, the evaluation metrics used in this study are defined formally below.

For a given label ℓ, let TP_ℓ, FP_ℓ, and FN_ℓ denote the number of true positives, false positives, and false negatives, respectively. Precision, recall, and F1 score for label ℓ are defined as in Equation (12).

$$\begin{aligned} \text{Precision}_\ell &= \frac{TP_\ell}{TP_\ell + FP_\ell}, \ \text{Recall}_\ell = \frac{TP_\ell}{TP_\ell + FN_\ell}, \\ \text{F1}_\ell &= \frac{2 \cdot \text{Precision}_\ell \cdot \text{Recall}_\ell}{\text{Precision}_\ell + \text{Recall}_\ell} \end{aligned} \tag{12}$$

The macro-averaged F1 score is computed as in Equation (13), where L is the number of modeled labels.

$$\text{Macro-F1} = \frac{1}{L} \sum_{\ell=1}^{L} \text{F1}_\ell \tag{13}$$

The weighted-averaged F1 score is defined as in Equation (14), where n_ℓ denotes the number of positive samples for the label ℓ.

$$\text{Weighted-F1} = \sum_{\ell=1}^{L} \frac{n_\ell}{\sum_{k=1}^{L} n_k} \text{F1}_\ell \tag{14}$$

Let $M^{(k)}$ represent the evaluation metric computed on the fold k, with $K = 5$. The cross-validated estimate of the metric M is given by Equation (15).

$$\widehat{M} = \frac{1}{K} \sum_{k=1}^{K} M^{(k)} \tag{15}$$

To compare the two best-performing supervised models, fold-level paired differences were analyzed. Normality of the paired metric differences was assessed using the Shapiro–Wilk test. When normality was not rejected, statistical significance was evaluated using a paired t-test; otherwise, a Wilcoxon signed-rank test was applied. For the paired t-test, let Equation (16) denote the difference in macro-F1 between Linear SVM and Logistic Regression on fold k.

$$d_k = M_{\text{SVM}}^{(k)} - M_{\text{LR}}^{(k)} \tag{16}$$

The test statistic is illustrated in Equation (17), where $\bar{d}$ is the mean of the paired differences and s_d is their standard deviation.

$$t = \frac{\bar{d}}{s_d/\sqrt{K}} \tag{17}$$

Effect size was quantified using Cohen's d as in Equation (18). As the effect size reflects fold-level consistency rather than absolute score magnitude, a large Cohen's d indicates stable directional superiority across folds rather than a dramatic increase in raw predictive performance.

$$d = \frac{\bar{d}}{s_d} \tag{18}$$

In addition, McNemar's test was applied to the out-of-fold exact-match predictions to assess whether the two classifiers differed systematically in their instance-level correctness. Let b and c denote the counts of discordant prediction pairs in the 2×2 contingency table as in Equation (19).

$$\chi^2 = \frac{(\mid b - c \mid -1)^2}{b + c} \tag{19}$$

3.1.11. Ablation design

To quantify the contribution of the preprocessing and representation choices, the strongest full model was re-estimated under four ablations: removal of metadata features, use of raw text without normalization, restriction to unigram TF–IDF, and title-only text input. Each ablation preserved the original cross-validation structure and metric set to ensure direct comparability. All derived outputs were written outside the raw data file. Fixed seeds, explicit preprocessing steps, and saved summaries were used to ensure the analysis can be rerun and audited.

4. Results

Results are presented in two stages: pre-model portfolio diagnostics and predictive model evaluation.

4.1. Pre-model portfolio diagnostics

These analyses serve as pre-model diagnostics, characterizing the statistical structure of the portfolio that the subsequent text-classification models attempt to recover. Table 2 represents the association between funding agency type and the prioritization of skill development, which was not only statistically significant (χ^2 = [67.46], p < .001) but also exhibited a strong effect size (Cramer's V = [0.327]). This indicates that the mandate-driven specialization of PMU-A is a robust characteristic of its funding portfolio rather than a marginal variation.

4.2. Distribution of socio-economic pain points

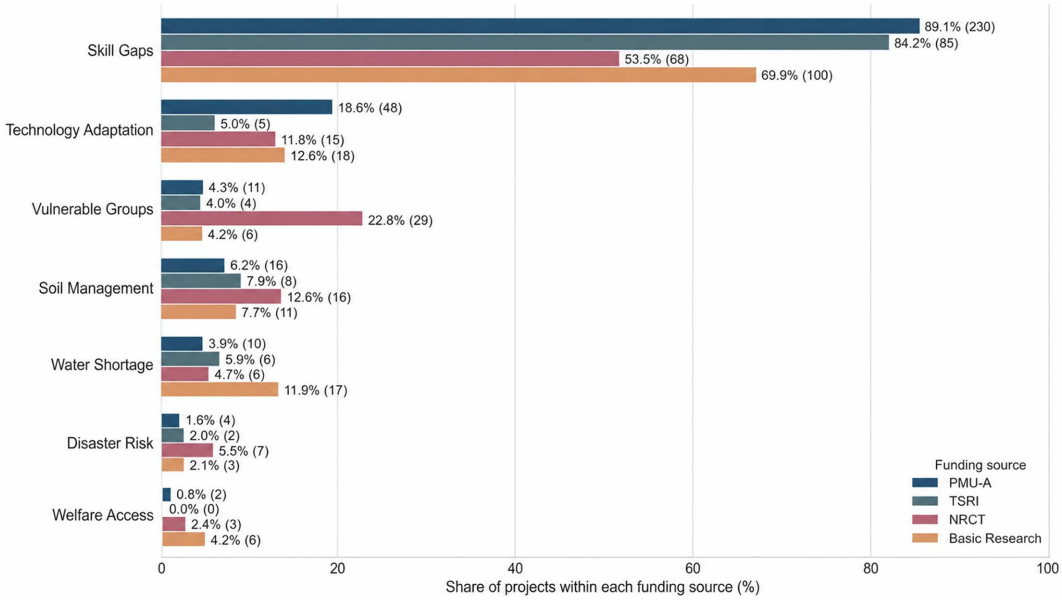
The distribution analysis of 629 research projects in Figure 1 reveals a distinct strategic specialization among the four agencies. PMU-A and Thailand Science Research and Innovation (TSRI) exhibit a heavy concentration on Skill Gaps, accounting for 89.1% and 84.2% of their portfolios, respectively. Conversely, National Research Councilof Thailand (NRCT) displays the most heterogeneous distribution, with significant investments in "Vulnerable

Table 2
Association between funding agency type and poverty-related portfolio priorities

Variable	Agency	*n* (%)	95% CI
Skill Gaps	PMU-A	230 (89.1)	[85.4–92.9]
	NRCT	68 (53.5)	[44.9–62.2]
	TSRI	85 (84.2)	[77.0–91.3]
	Basic	100 (69.9)	[62.4–77.4]
Vulnerable Groups	PMU-A	11 (4.3)	[1.8–6.7]
	NRCT	29 (22.8)	[15.5–30.1]
	TSRI	4 (4.0)	[0.2–7.8]
	Basic	6 (4.2)	[0.9–7.5]
Water Scarcity	PMU-A	10 (3.9)	[1.5–6.2]
	NRCT	6 (4.7)	[1.0–8.4]
	TSRI	6 (5.9)	[1.3–10.5]
	Basic	17 (11.9)	[6.6–17.2]
Technology Adaptation	PMU-A	49 (19.0)	[14.2–23.8]
	NRCT	15 (11.8)	[6.2–17.4]
	TSRI	5 (5.0)	[0.7–9.2]
	Basic	18 (12.6)	[7.2–18.0]
Soil Management	PMU-A	16 (6.2)	[3.3–9.1]
	NRCT	16 (12.6)	[6.8–18.4]
	TSRI	8 (7.9)	[2.6–13.2]
	Basic	11 (7.7)	[3.3–12.1]

Note: Values are reported as *n* (%) with 95% confidence intervals (CI). Sample sizes: PMU-A (N = 258), NRCT (N = 127), TSRI (N = 101), Basic Research (N = 143).

Figure 1
Strategic focus areas (PMU-A vs other agencies)



Groups" (22.8%) and "Soil Management" (12.6%). These data suggest that PMU-A serves as a specialized arm for productivity enhancement. In contrast, NRCT maintains a broader mandate that addresses the multidimensional nature of poverty, consistent with ABD, inclusive innovation, and regional innovation-system perspectives that emphasize locally differentiated development priorities [2–5].

The association between funding agency types and the prioritization of Skill Gaps was found to be statistically significant ($\chi^2 = 67.46$, $p < 0.001$). Furthermore, the analysis yielded a Cramer's V of 0.327, indicating a moderate-to-strong effect size. This confirms that the observed specialization of PMU-A in capacity-building interventions is not a result of random variation but reflects a deliberate and robust strategic mandate. PMU-A demonstrated an overwhelming focus on skill development, accounting for 89.1% of its portfolio, with a high level of precision (95% CI [85.4%, 92.9%]). In stark contrast, national agencies such as NRCT placed significantly less emphasis on the same pillar (53.5%; 95% CI [44.9%, 62.2%]). The fact that these confidence intervals do not overlap provides strong statistical evidence of a clear strategic bifurcation in the Thai research ecosystem, in which area-based units specialize in economic empowerment while national agencies maintain broader social welfare mandates, a pattern that aligns with prior arguments that local development interventions are shaped by institutional mandates, territorial needs, and governance structures [2, 4, 16].

4.3. Strategic pillar comparison across funding bodies

Figure 2 aggregates the pain points into four strategic pillars, which underscore a functional division of labor. The "Human and Productivity Capital" pillar is the primary focus for all agencies, yet its dominance is most absolute in PMU-A (93.8%). Notably, the "Natural Capital" pillar, comprising water and soil management, receives comparatively lower attention from area-based funding (6.2%) than from Basic Research (14.7%) and NRCT (19.7%). This gap indicates a potential area for future policy alignment to ensure that skill-based interventions are supported by adequate natural resource management because area-based and inclusive innovation approaches emphasize that capability development must be coordinated with local environmental, institutional, and livelihood constraints [2–5].

4.4. Portfolio composition and institutional mandates

Figure 3 illustrates the internal allocation of research projects across the four funding agencies, categorized by strategic pillars. The results highlight the mandate-driven nature of each organization: PMU-A and TSRI exhibit highly specialized, "Productivity-Centric" portfolios, with over 80% of their resources dedicated to Human and Productivity Capital. In contrast, the NRCT displays a more "Balanced" distribution, with a significant portion (23.9%) of its projects addressing social welfare and the needs of vulnerable groups. These distributions confirm a functional division of labor within the Thai research ecosystem and support the view that area-based innovation systems are shaped by interactions among funding agencies, universities, local actors, and policy mandates rather than by project-level activity alone [2, 4, 16].

4.5. Inter-label correlation of poverty pain points

The correlation heatmap in Figure 4 provides a concise overview of the interrelationships among the poverty-related indicators, thereby highlighting important socio-economic synergies and patterns of co-occurrence within the studied communities. The most pronounced positive associations were observed between Welfare Access and High Dependency, and between Water Shortage and Disaster Risk, suggesting that these dimensions of deprivation tend to cluster together within the portfolio. By contrast, several associations involving Lack Skills were negative, indicating that skill-related constraints may follow a distinct pattern relative to other poverty dimensions and therefore warrant separate analytical and policy consideration. These correlations provide a statistical basis for designing cross-functional research calls that address multiple interlocking pain

Figure 2
Strategic pillar distribution across funding bodies

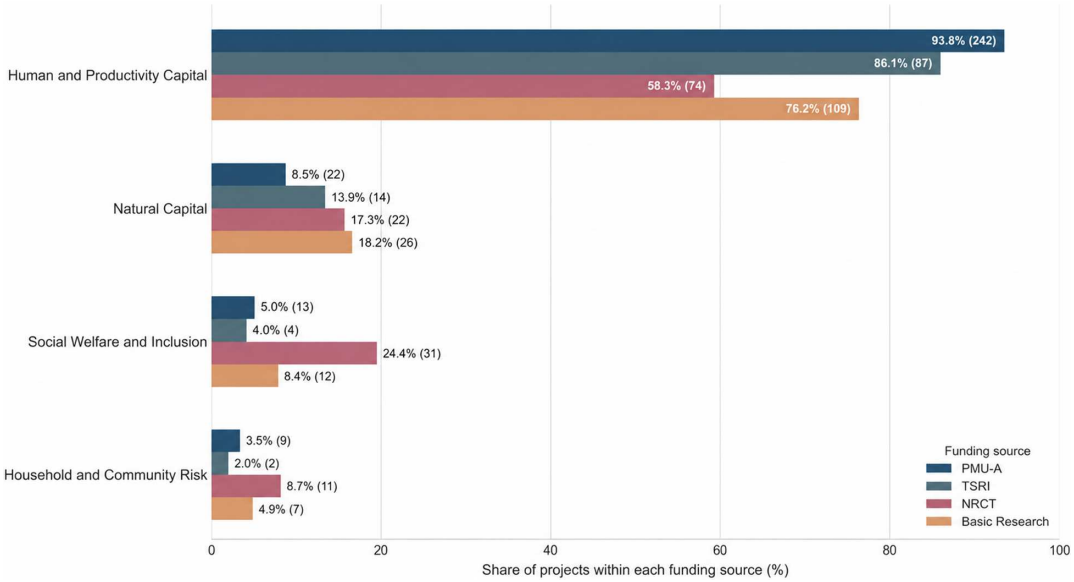
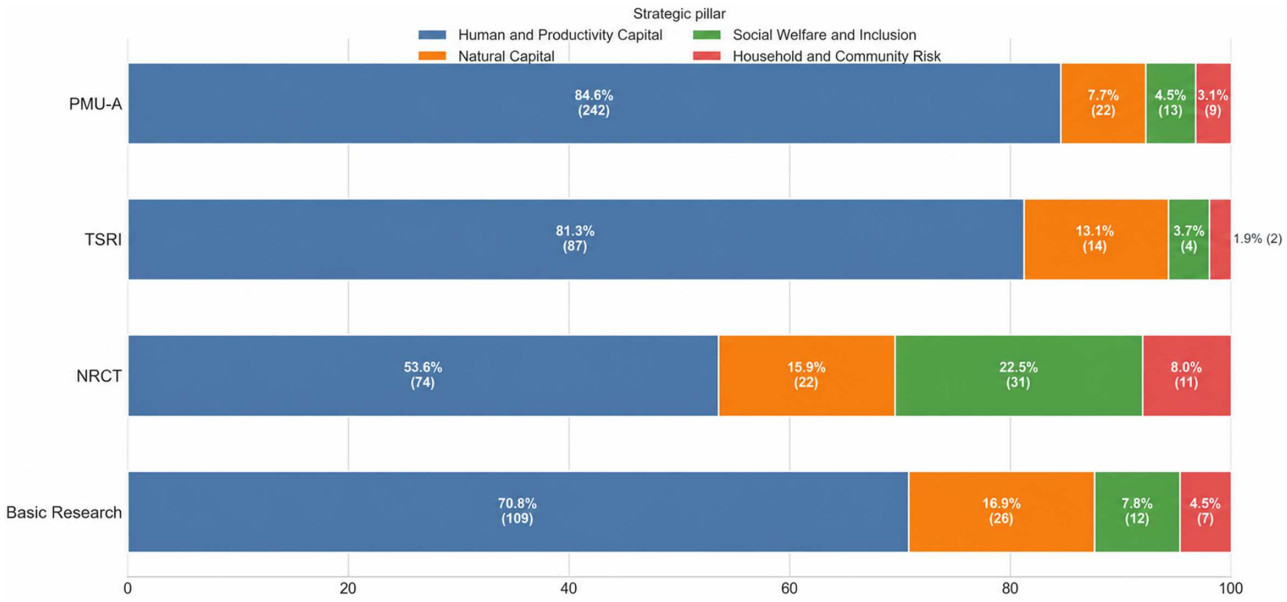


Figure 3
Internal portfolio composition and strategic mandates by funding agency



points simultaneously, consistent with inclusive innovation and community-based development perspectives that treat poverty as a multidimensional problem requiring coordinated local responses [1, 3, 5].

4.6. Model evaluation

To extend the descriptive and inferential portfolio analysis, the manuscript also evaluated whether poverty-oriented project categories could be detected automatically from project text. Table 3 shows the processed dataset, which contains 629 project records, 22 columns, and no missing values. Mean label cardinality is 1.203, and mean text length is 9.143 tokens, which confirms that the corpus is both short and multi-label. The funding portfolio is concentrated in a few sources: PMU-A, Basic Research, NRCT, and TSRI. The class structure is highly imbalanced, with Lack_Skills far more common than the rarer labels.

This profile shows that the empirical task is low-resource, short-text, and label-imbalanced. That combination makes sparse linear methods more defensible than deep sequence models and explains why macro-F1 is a more meaningful criterion than raw accuracy in this setting [18, 21–24].

Figure 5 summarizes label prevalence across the full target set. The distribution is strongly skewed toward a few common labels, especially Lack_Skills. Only seven labels met the support threshold for supervised modeling, which is why the article emphasizes macro-averaged evaluation and descriptive reporting for the rarer categories. The visual also supports the claim that label sparsity, rather than model complexity, is the primary constraint on broader inference, which is consistent with prior

Figure 4
Inter-label correlation structure among poverty-related pain points

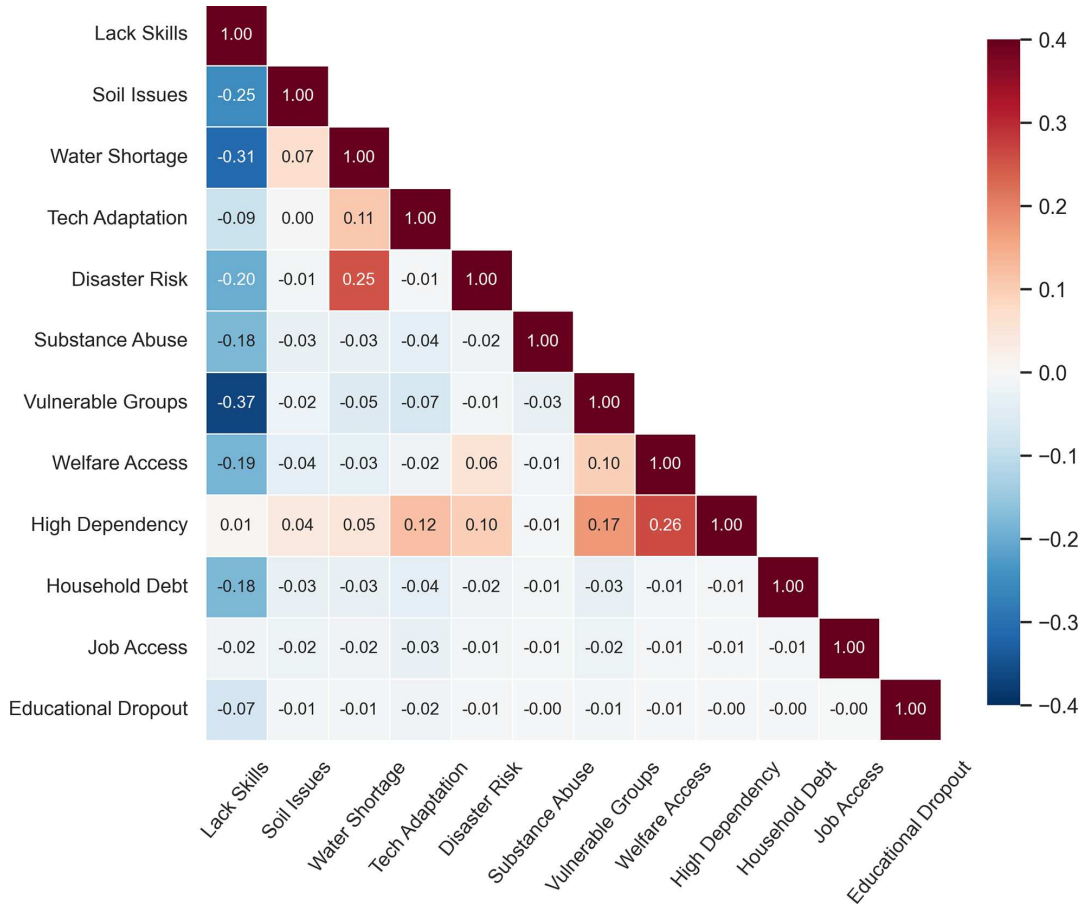
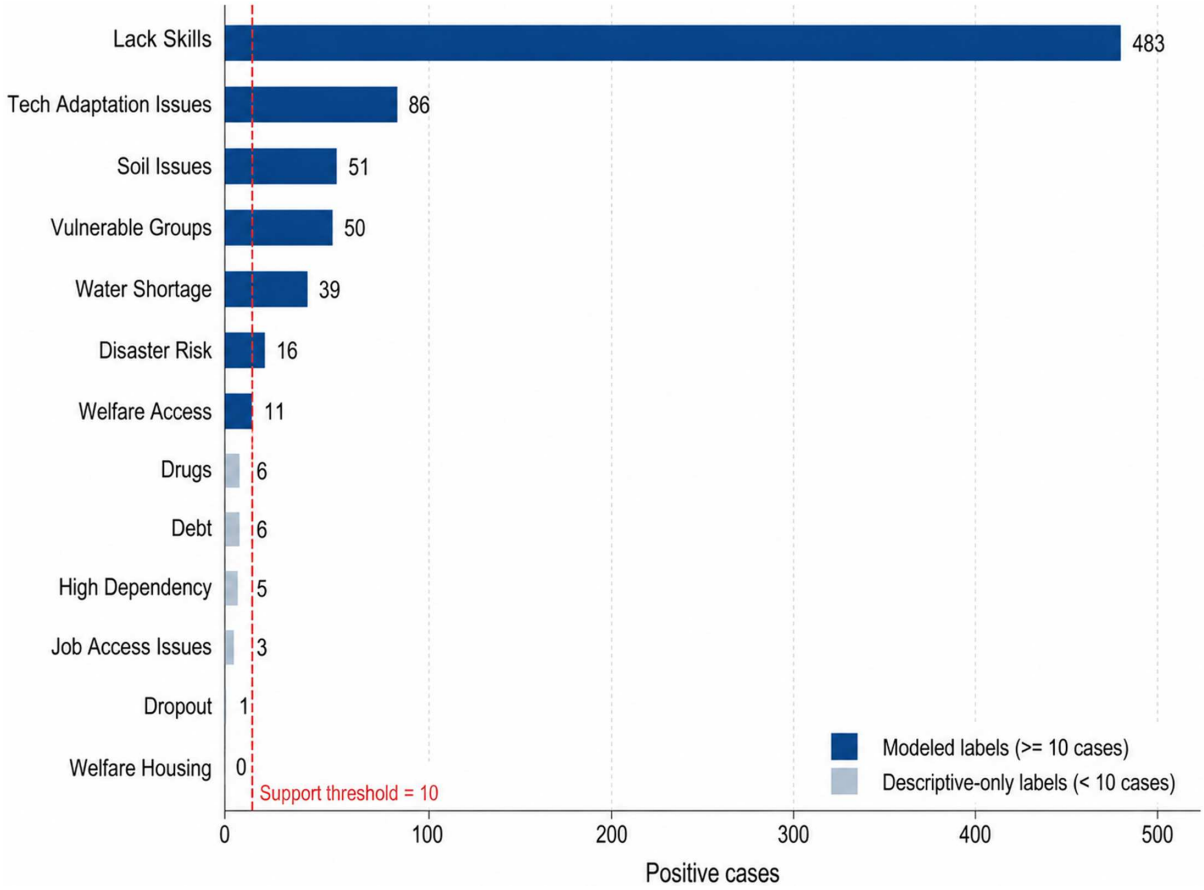


Table 3
Dataset characteristics

Metric	Value
Projects	629
Columns	22
Observed years	2020, 2021, 2024
Minimum text tokens per project	3
Median text tokens per project	9
Maximum text tokens per project	28
Mean labels per project	1.203
Mean text tokens per project	9.143
Funding sources	PMU-A: 258; Basic Research: 143; NRCT: 127; TSRI: 101
Labels retained for supervised modeling	7
Labels retained descriptively only	6
Constant labels	1 (Welfare_Housing)

Note: Dataset characteristics describe the analytical file used for multi-label text classification. Text-length statistics should be interpreted as evidence of a short-text, low-resource classification setting. The observed years are discontinuous because records for 2022 and 2023 were not available in the analytical file.

Figure 5
Label distribution for poverty-related indicators



work showing that imbalanced multi-label settings require careful metric selection and cautious interpretation of minority-label performance [6–8, 18, 21].

Figure 6 shows the funding-source composition of the portfolio. PMU-A is the largest source in the portfolio, followed by Basic Research, NRCT, and TSRI. Please note that this is descriptive portfolio information only. It should not be interpreted as evidence that a particular funding source leads to better poverty outcomes or different label structures because the dataset does not identify a causal mechanism of assignment.

4.7. Main analytical findings

Table 4 reports the mean cross-validated performance and fold-level standard deviation for all evaluated models under the same five-fold protocol. Linear SVM achieved the strongest macro-F1 performance (0.5391 ± 0.0583), followed by Logistic Regression (0.5114 ± 0.0598), indicating that the two sparse linear models provided the best balance between minority-label recovery and overall predictive stability. Linear SVM also achieved the highest weighted-F1 score (0.7798 ± 0.0171), while its accuracy (0.6343 ± 0.0297) remained comparable to the tree-based models. This pattern supports selecting Linear SVM as the main model, as its advantage is clearest on macro-F1, the most appropriate metric for the imbalanced multi-label setting [21–24].

The comparison also shows why accuracy alone is insufficient. XGBoost and Random Forest achieved slightly higher or comparable accuracy values (0.6408 ± 0.0230 and 0.6408 ± 0.0279, respectively), but their macro-F1 scores were substantially lower than those of Linear SVM and Logistic Regression. This indicates that the tree-based models performed relatively well on dominant label patterns but were less effective at recovering minority labels. Similarly, the majority baseline achieved relatively high accuracy (0.6280 ± 0.0082) but extremely low macro-F1 (0.1241 ± 0.0014), confirming that high accuracy can be misleading when the label distribution is highly imbalanced [22–24].

The keyword heuristic produced a moderate macro-F1 score (0.4468 ± 0.0831), suggesting that transparent lexical rules captured some poverty-related signal. However, its low weighted-F1 (0.2891 ± 0.0806) and accuracy (0.1304 ± 0.0440) indicate poor overall generalization and many incorrect predictions. Sparse linear classifiers, especially Linear SVM, provide the most defensible trade-off between interpretability, minority-label sensitivity, and cross-validated predictive performance in this short-text, label-imbalanced portfolio-audit task [6–8, 17, 18, 21, 24].

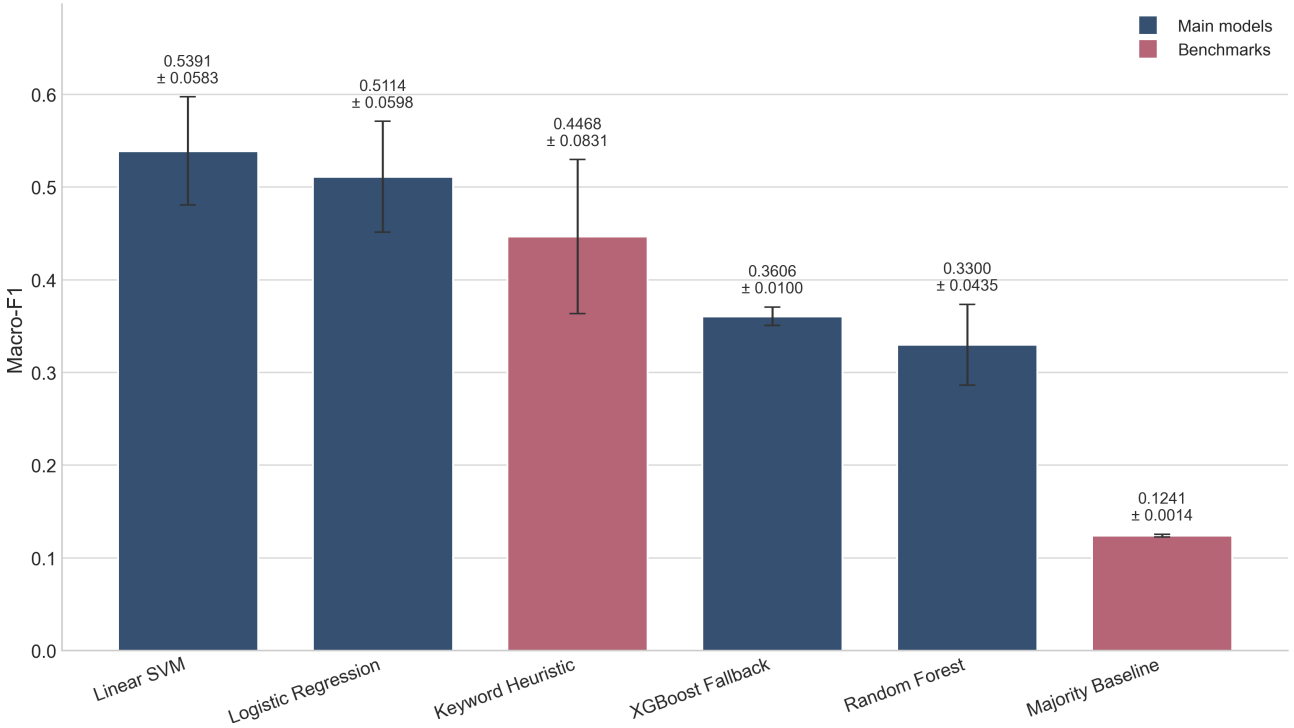
Figure 6 visualizes the model comparison. The figure makes the same pattern immediately visible. The two linear models lead the pack, while the tree-based models trail behind. The keyword heuristic is informative as a benchmark because it captures some lexical signals, but it still lags behind the supervised linear models. This supports the interpretation that the project descriptions contain learnable structure beyond simple label priors, while also confirming the value of comparing supervised models with transparent baselines in low-resource text classification tasks [8, 17, 18].

Table 4
Comparison of model characteristics

Model	Type	Macro-F1 mean ± SD	Weighted-F1 mean ± SD	Accuracy mean ± SD	Note
Linear SVM	Main model	0.5391 ± 0.0583	0.7798 ± 0.0171	0.6343 ± 0.0297	Best overall macro-F1
Logistic Regression	Main model	0.5114 ± 0.0598	0.7582 ± 0.0143	0.5883 ± 0.0451	Second-best macro-F1
XGBoost Fallback	Main model	0.3606 ± 0.0100	0.7062 ± 0.0199	0.6408 ± 0.0230	High-capacity fallback
Random Forest	Main model	0.3300 ± 0.0435	0.6939 ± 0.0302	0.6408 ± 0.0279	Weakest trained model
Keyword Heuristic	Baseline	0.4468 ± 0.0831	0.2891 ± 0.0806	0.1304 ± 0.0440	Transparent rule baseline
Majority Baseline	Baseline	0.1241 ± 0.0014	0.5703 ± 0.0189	0.6280 ± 0.0082	The weakest evaluated system

Note: Values are reported as cross-validation mean ± standard deviation. Macro-F1 is the primary evaluation metric because the task is multi-label and highly imbalanced. Weighted-F1 and accuracy are reported as secondary metrics because they may be dominated by high-support labels.

Figure 6
Model comparison by macro-F1



4.8. Ablation and robustness

Table 5 reports the ablation results for the Linear SVM model under the same cross-validation protocol. The unigram-only specification achieved the highest macro-F1 score (0.5445), slightly exceeding the full model (0.5391), whereas the title-only configuration produced the largest performance decline (macro-F1 = 0.4385). This comparison shows that the innovation-description text contributes essential discriminative signal, while additional bigram and metadata features do not provide a net performance gain in this short-text setting, consistent with prior evidence that representation choices in text classification should be evaluated empirically rather than assumed to improve performance [6, 7, 18].

The no-metadata ablation produced a small decline in macro-F1 (0.5294), suggesting that administrative variables contribute only marginally relative to the project text. The no-text-normalization setting produced identical performance to the full model, indicating that normalization choices had limited impact under the sparse TF–IDF representation. The ablation evidence supports a parsimonious unigram-based representation. It shows that title-only text is insufficient for reliable classification, reinforcing the importance of evaluating preprocessing and feature contributions through controlled model comparison [6, 18, 25].

Table 5
Ablation study

Ablation	Macro-F1	Weighted-F1	Delta vs full model
Unigram only	0.5445	0.7753	+0.0054
Full model	0.5391	0.7798	0.0000
No metadata	0.5294	0.7743	–0.0097
No text normalization	0.5391	0.7798	0.0000
Title only	0.4385	0.7131	–0.1007

Note: Delta values are computed relative to the full Linear SVM model. The unigram-only result indicates that the full bigram and metadata representation did not provide a net performance gain. In contrast, the title-only result shows that innovation-description text contributes essential signal.

The unigram-only advantage should not be interpreted as evidence that all bigrams are useless. A specific bigram such as "poverty data" may still be highly discriminative within the full model, even if the full bigram feature space introduces sparsity, noise, or fold-level instability. Therefore, the ablation result and feature-importance result are not contradictory: ablation evaluates the net contribution of an entire feature family, whereas feature importance identifies influential terms within a fitted model [9–11, 25].

Figure 7 visualizes the ablation comparison. The figure reinforces the table. A simpler unigram design is slightly better than the full configuration, whereas removing the innovation text results in a clear loss. The practical implication is that the article should not overclaim the value of administrative metadata. In this dataset, the strongest signal is still the project language itself.

4.9. Statistical validation

Table 6 reports the inferential comparison between the two best-performing models, Linear SVM and Logistic Regression, using paired fold-level evaluation. Normality of the paired macro-F1 differences is not rejected by the Shapiro–Wilk test ($p = 0.6939$), supporting the use of a paired t-test for inference. The paired t-test yields a statistically significant result ($t = 7.0524$, $p = 0.0021$), indicating that the mean macro-F1 of Linear SVM is reliably higher than that of Logistic Regression under the shared cross-validation protocol. The estimated effect size is large (Cohen's $d = 3.1539$), reflecting strong consistency in the direction of the fold-level differences rather than a large absolute performance gap. The 95% confidence interval for the mean macro-F1 difference [0.0168, 0.0387] further excludes zero, reinforcing this conclusion.

McNemar's test on out-of-fold exact-match predictions is also statistically significant ($\chi^2 = 9.0115$, $p = 0.0027$), indicating systematic differences in instance-level correctness between the two classifiers. Together, these tests show that the observed advantage of Linear SVM over Logistic Regression is not attributable to random variation induced by cross-validation partitioning. Instead, it reflects a statistically reliable difference within this dataset and evaluation design.

Figure 7
Ablation performance for the Linear SVM

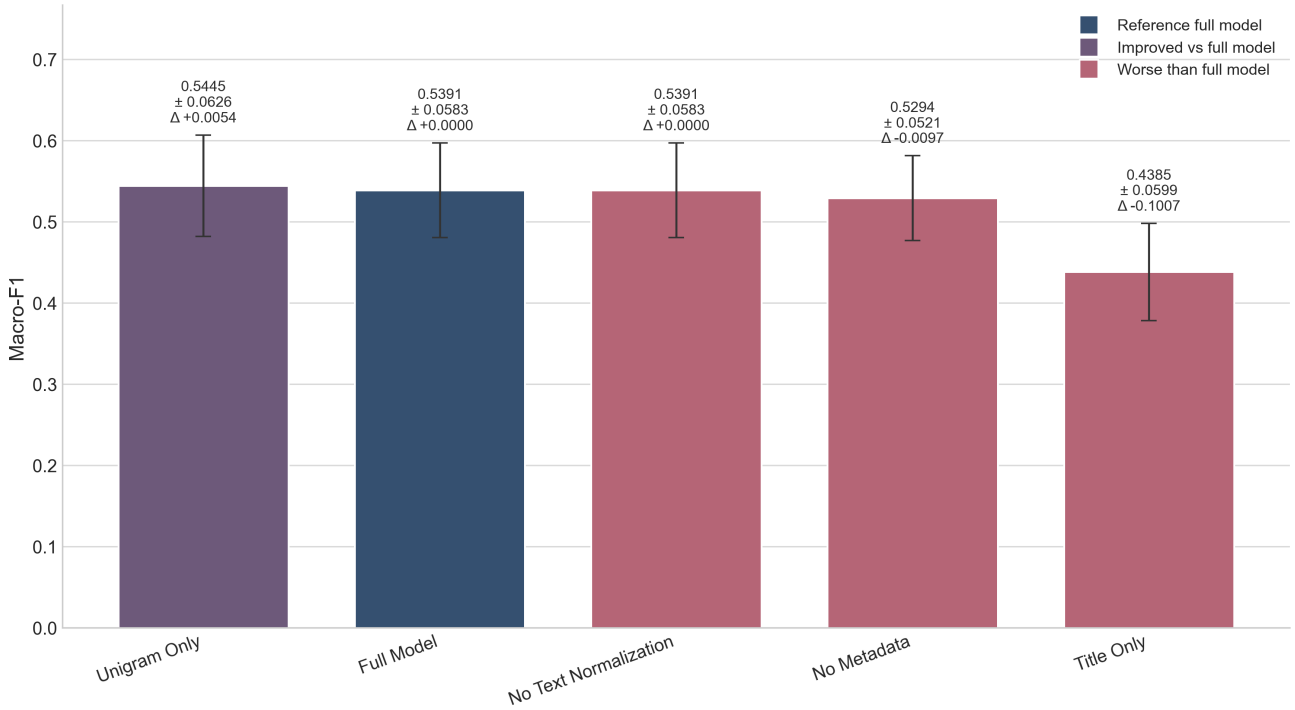


Table 6
Comparison of Linear SVM and Logistic Regression

Test	Comparison	Statistic	*p*-value	Effect/interval
Shapiro–Wilk	Paired fold differences	0.9439	0.6939	Normality not rejected
Paired *t*-test	Linear SVM vs Logistic Regression	7.0524	0.0021	Cohen's d = 3.1539; 95% CI = [0.0168, 0.0387]
McNemar	Exact-match outcomes	9.0115	0.0027	Contingency table = [[341, 58], [29, 201]]

Note: Statistical tests are based on five cross-validation folds and out-of-fold exact-match predictions. Cohen's d should be interpreted as a fold-level consistency indicator rather than as evidence of a large absolute performance difference.

The inferential results remain confined to classifier comparison and do not support substantive conclusions about the effect of research funding on poverty alleviation. Predictively, Linear SVM is the strongest model under the chosen metrics and validation protocol. Inferentially, its advantage over Logistic Regression is statistically supported within this sample. Consistent with established guidance, predictive performance and inferential validation are interpreted separately from any causal claims about policy impact [26, 27].

The paired statistical comparison should also be interpreted cautiously because it is based on five cross-validation folds. Although the Linear SVM consistently outperformed Logistic Regression across folds and the paired test was statistically significant, the number of paired observations is small. Therefore, Cohen's d should be understood as a fold-level consistency indicator rather than as evidence of a large absolute improvement in predictive performance. The absolute macro-F1 difference remains modest. For this reason, the revised interpretation emphasizes both statistical reliability and practical moderation: Linear SVM is the strongest model evaluated, but its advantage should be validated on larger, externally held-out funding portfolios [26, 27].

Table 7 provides additional multi-label diagnostics for the seven supervised poverty-related labels. These results complement the aggregate macro-F1 and weighted-F1 scores by showing how Linear SVM performed across individual labels with different levels of positive-case support. The table is important because aggregate metrics can obscure whether performance is distributed across labels or driven mainly by the most frequent category, a known concern in imbalanced and multi-label classification evaluation [21–24].

4.10. Visual evidence and explainability

Figure 8 shows the confusion matrix for the best model, Linear SVM. The matrix shows the expected behavior of a model trained on imbalanced multi-label data: frequent categories are recognized more reliably than rare ones. That pattern is consistent with the macro-F1 results. It reinforces the need to interpret the model as a screening tool rather than as a perfect classifier, especially in imbalanced multi-label settings where frequent categories are typically easier to recover than rare ones [21, 22, 24].

Since the results indicate that the Linear SVM achieves the highest performance, Table 8 ranks the top TF–IDF features contributing to the Linear SVM predictions. These features provide an interpretable view of the lexical signals the model uses and help determine whether the classifier learned policy-relevant thematic cues from project descriptions rather than relying on opaque or arbitrary patterns [9–11, 13, 14]. The table ranks the top TF–IDF features by their contribution to model predictions, with importance scores ranging from 0.611 to 1.149.

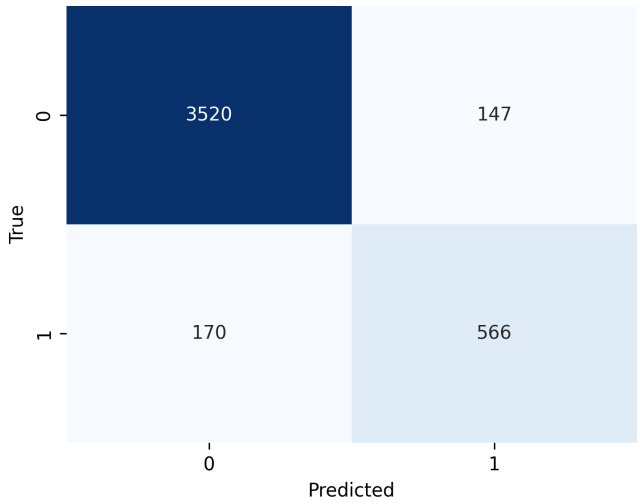
The most influential feature, *"poverty data,"* has the highest importance (1.149) and a strong positive weight (1.149), followed by "soil" (importance = 1.040, weight = 0.347) and "water" (importance = 0.938, weight = 0.280). Environmental and risk-related terms such as *"disaster"* (0.788, 0.339) and *"risk"* (0.681, 0.167) also contribute substantially. Notably, *"data"* (0.714) and *"poverty"* (0.611) exhibit fully aligned signed weights, indicating

Table 7
Additional multi-label performance diagnostics for the supervised labels

Label	Positive cases	F1-score	Precision	Recall
Lack_Skills	483	0.9236	0.9343	0.9130
Soil_Issues	51	0.5319	0.5814	0.4902
Water_Shortage	39	0.6494	0.6579	0.6410
Tech_Adaptation_Issues	86	0.4181	0.4066	0.4302
Disaster_Risk	16	0.3846	0.5000	0.3125
Vulnerable_Groups	50	0.5545	0.5490	0.5600
Welfare_Access	11	0.5263	0.6250	0.4545
Overall Hamming Loss	–	0.0720	–	–

Note: Per-label precision, recall, F1-score, and overall Hamming loss were computed from the same out-of-fold prediction outputs used for the main Linear SVM evaluation. These diagnostics show whether the aggregate macro-F1 reflects performance across retained poverty-related labels or is mainly driven by high-support labels. Labels with small positive-case support, especially Disaster_Risk and Welfare_Access, should be interpreted cautiously because a small number of errors can strongly affect their precision, recall, and F1-score.

Figure 8
Confusion matrix for Linear SVM



a direct and consistent positive influence on classification outcomes. In contrast, features such as *"elderly"* show relatively low impact (importance = 0.637, weight = 0.012), suggesting a weaker or more context-dependent role. The results indicate that high-importance features are predominantly associated with socio-economic vulnerability and environmental risk, with the magnitudes of their contributions reflected in their signed weights; however, such weights should be interpreted as model associations rather than causal evidence [9–11, 15].

Figure 9 visualizes the feature importance for Linear SVM. The dominant features are lexically intuitive and policy-relevant: poverty data, soil, water, disaster, Geographic Information System (GIS) data risk, elderly, low, and poverty. This supports the claim that the model is learning interpretable thematic cues from the project descriptions rather than opaque latent representations. At the same time, the presence of these obvious terms also implies that the model is best understood as an explainable screening device, not a substitute for substantive domain judgment [9–11].

5. Discussion

5.1. Interpretation of the predictive findings

The results show that explainable classical machine-learning methods can recover meaningful thematic structure from a low-resource research funding portfolio. Linear SVM achieved the strongest macro-F1 among the evaluated models, while Logistic Regression remained the second-best supervised model. This pattern suggests that sparse linear classifiers are well suited to short administrative project texts when labels are imbalanced and interpretability is important [8, 17, 18, 21, 24]. The finding should not be read as a universal rejection of deep learning. Rather, it shows that, under the observed data conditions, transparent sparse models provide a defensible balance between predictive performance, reproducibility, and policy-facing interpretability [9–11].

The moderate macro-F1 score is also important. A macro-F1 of 0.5391 indicates useful but imperfect screening capability. The model can support portfolio review by identifying likely poverty-related thematic categories, but it should not replace expert judgment, agency review, or manual validation. In practice, the model is most useful as a first-pass screening tool: it can flag projects that appear strongly aligned with particular poverty dimensions, identify potential under-labeled records, and help funding managers inspect thematic concentration across agencies. However, final interpretation should remain human-reviewed because short project texts may not fully capture project objectives, implementation design, or real-world impact [9–11, 13, 15].

5.2. Theoretical implications for area-based development and inclusive innovation

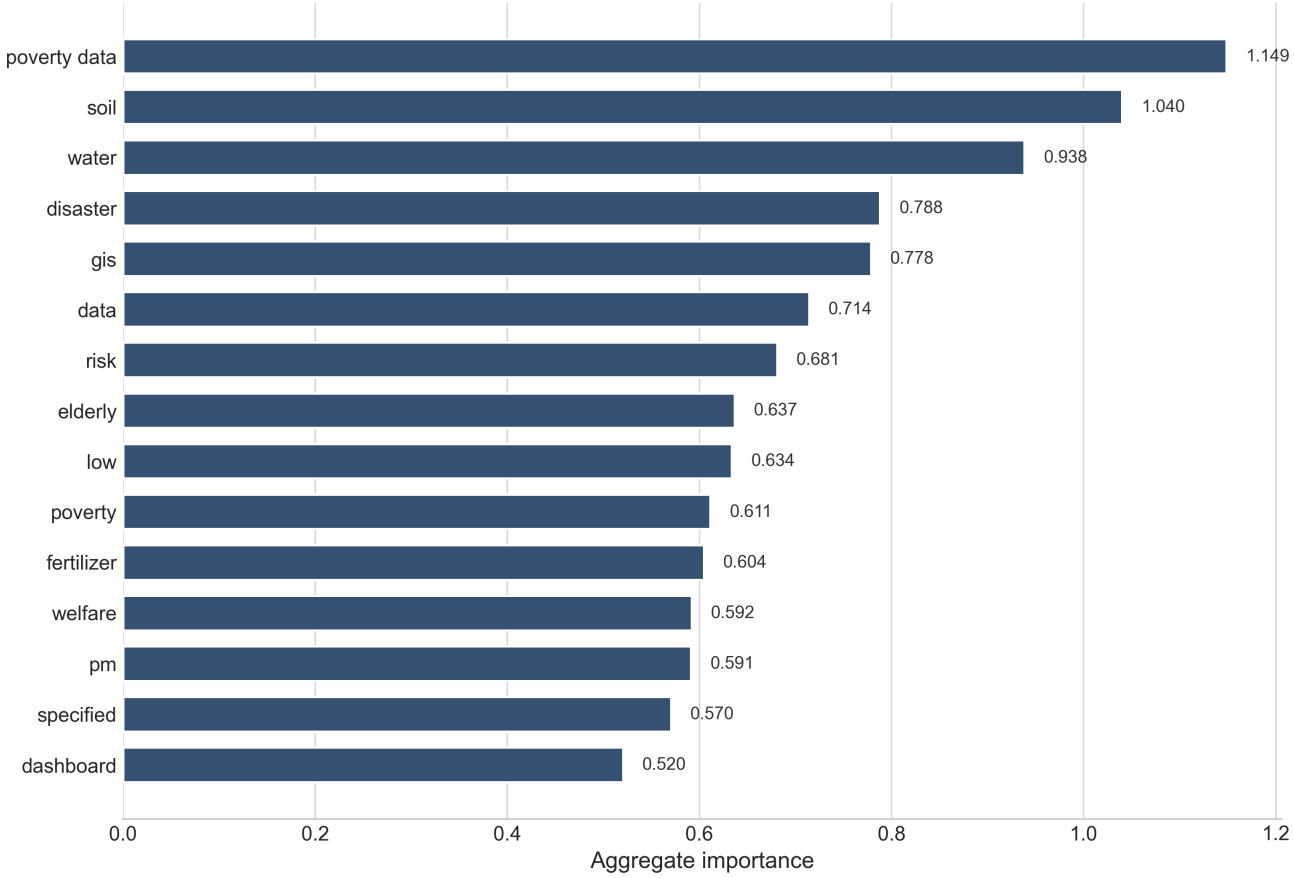
The findings contribute to ABD by showing that project language can be treated as a measurable expression of place-based policy priorities. The high concentration of skill-related terms in some agency portfolios suggests that research funding is not random in theme; rather, it reflects institutional mandates and strategic assumptions about how poverty should be addressed. This is consistent with ABD theory, which emphasizes that

Table 8
Top 10 features and their importance scores

Rank	Feature	Importance	Signed weight
1	tfidf__poverty data	1.149	1.149
2	tfidf__soil	1.040	0.347
3	tfidf__water	0.938	0.280
4	tfidf__disaster	0.788	0.339
5	tfidf__gis	0.778	0.113
6	tfidf__data	0.714	0.714
7	tfidf__risk	0.681	0.167
8	tfidf__elderly	0.637	0.012
9	tfidf__low	0.634	0.128
10	tfidf__poverty	0.611	0.611

Note: Feature-importance values indicate the relative contribution of TF–IDF features to Linear SVM predictions. Signed weights indicate the direction and magnitude of model association within the fitted classifier. These values support model explainability and lexical transparency, but they should not be interpreted as causal effects, project impact, or evidence that the named terms alone define poverty-alleviation outcomes.

Figure 9
Feature importance for the best model



institutional roles, local capabilities, and contextual needs shape local development interventions [2, 4, 16].

The results also speak to Regional Innovation Systems and inclusive innovation. In a regional innovation system, universities, funding agencies, community enterprises, local governments, and knowledge intermediaries should interact to build capability and address local constraints. The portfolio patterns observed in this study suggest that some agencies emphasize human and productivity capital, while others maintain broader welfare or natural resource orientations. This division of labor can be beneficial when coordinated. Still, it may also create gaps if natural capital, access to welfare, or vulnerable groups receive insufficient attention within area-based portfolios [3–5].

The study also responds to the "local trap" critique. Area-based funding should not assume that local targeting is automatically more democratic, inclusive, or effective. A project may be locally framed yet still neglect important dimensions of poverty. Text-based portfolio auditing can therefore help funding agencies assess whether local research portfolios are genuinely balanced across human capital, natural capital, social welfare, and technology-adaptation needs [16].

5.3. Methodological implications

The comparison between feature settings shows that simpler representations can be more robust than more complex ones in low-resource text classification. The unigram-only model slightly outperformed the full model, suggesting that additional bigrams and metadata did not improve overall generalization. However, the feature-importance results indicate that individual bigrams, such as "poverty data," may still be highly informative. This distinction is important: ablation evaluates the net contribution of an entire feature family, while feature importance identifies influential individual terms inside a fitted model [6, 7, 18, 25].

The statistical validation supports Linear SVM as the strongest model evaluated, but the small number of cross-validation folds warrants caution. The large Cohen's d indicates stable fold-level superiority, not a dramatic absolute gain. The revised interpretation therefore treats Linear SVM as the best model within this dataset and evaluation design, while recommending external validation before operational deployment [26, 27].

5.4. Practical implications for Thai funding agencies

The findings can support research-governance practice in several ways. First, funding agencies can use text-based auditing to monitor whether poverty-related portfolios are concentrated too heavily in a small number of themes, such as skill development, while under-representing natural resource constraints, welfare access, or vulnerable groups. Second, agencies can improve data-entry standards by requiring clearer project descriptions, explicit poverty-dimension tags, and consistent innovation descriptors. Third, portfolio dashboards can incorporate explainable text classification outputs to support periodic monitoring of thematic balance across funding calls. Fourth, area-based funding programs can use the results to design more integrated calls that connect skill development with water, soil, disaster, welfare,

and technology-adaptation issues, consistent with inclusive innovation and ABD perspectives that emphasize locally embedded, multidimensional interventions [2–5].

For PMU-A and similar area-based agencies, the results suggest that productivity-oriented and human-capital interventions should be complemented by stronger attention to natural capital, especially soil and water constraints, where these are central to local poverty. For national agencies, the framework can help determine whether broader welfare and vulnerable-group mandates are consistently reflected in the language of funded projects. For universities and research administrators, the framework provides a reproducible method for reviewing whether project portfolios align with institutional missions, local development plans, and poverty-alleviation objectives.

The study offers a practical screening workflow for funding agencies, university administrators, and research evaluators. First, it can help summarize large project portfolios into interpretable poverty-related categories. Second, it can highlight which categories are sufficiently supported for automated analysis and which remain too sparse for stable modeling. Third, it provides a reproducible way to compare text-only and hybrid specifications without overfitting the results. This kind of screening workflow aligns with the policy-facing use of interpretable models in accountability settings [9–11, 13, 14].

5.5. Limitations

Several limitations should be acknowledged. First, the dataset contains 629 project records, which are useful for portfolio auditing but still limited for high-capacity machine-learning models. Second, the text fields are short, with project descriptions often containing only compact administrative language. This restricts semantic richness and makes the task sensitive to lexical cues [7, 8]. Third, the label distribution is highly imbalanced, and several poverty-related labels were too rare for stable supervised learning, a recognized challenge in multi-label classification [21–24]. Fourth, the dataset covers 2020, 2021, and 2024, but not 2022 or 2023. This discontinuity impedes robust longitudinal interpretation and may limit the generalizability of temporal patterns. Fifth, the analysis is based on project text and administrative labels rather than measured poverty outcomes. Therefore, the study cannot determine whether funded projects reduce poverty. Sixth, the One-vs-Rest framework does not directly model label dependency, even though label co-occurrence is substantively meaningful [19, 20]. Finally, the statistical comparison is based on five cross-validation folds and should be validated using larger external portfolios [26, 27].

5.6. Future research

Future research should expand the dataset to include additional funding years and agencies, including the missing 2022–2023 period, where available. Larger datasets would enable stronger comparisons among TF–IDF baselines, transformer embeddings, few-shot prompting, and parameter-efficient fine-tuning [7, 12, 17, 18]. In addition, future work should also evaluate multi-label methods that directly model label dependence, including classifier chains, label-powerset approaches, and neural multi-label architectures [19, 20, 24]. Another priority is to connect project-text classification with implementation outcomes, beneficiary data, and community-level poverty indicators. Such integration would allow future studies to move from portfolio screening toward more robust evaluation of research funding alignment, implementation quality, and social impact [1–3, 5].

The framework should also be validated using larger, more continuous datasets, including additional funding years and external agency portfolios. Funding agencies should improve the quality and consistency of project descriptions, require clearer poverty-dimension tagging, and use explainable text analytics as part of regular portfolio-monitoring workflows. Future studies should also compare sparse lexical baselines with transformer-based methods and evaluate multi-label approaches that model label dependency more directly [12, 17, 19, 20, 24].

6. Conclusion

This study developed and evaluated an explainable multi-label text classification workflow to audit a Thai area-based research funding portfolio. Using 629 project records, the analysis showed that poverty-related project categories can be predicted from short project text with moderate but useful accuracy under low-resource conditions. Linear SVM achieved the strongest macro-F1 among the evaluated models, while Logistic Regression remained a competitive and interpretable alternative. Ablation analysis indicated that unigram-only text slightly outperformed the full feature set, while title-only modeling substantially reduced performance. Feature-importance analysis showed that the model relied on interpretable lexical cues related to poverty, soil, water, disaster, and data. The study therefore contributes a reproducible, policy-facing framework for portfolio screening, thematic monitoring, and research-governance review, consistent with prior work on short-text classification, imbalanced multi-label evaluation, and interpretable machine learning [18, 21–24].

Acknowledgments

The data utilized in this study were sourced from the NRIIS. The authors gratefully acknowledge the support and cooperation of the Center for SDG Research and Support (SDG Move), Thammasat University, in facilitating the research process and the overall conduct of this study.

Funding Support

This research was financially supported by the TSRI, fiscal year 2025.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available on GitHub at https://github.com/pjarupunphol/Datasets/blob/main/Research_Funding_Data.zip.

Author Contribution Statement

Wichidtra Sudjarid: Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Project administration, Funding acquisition. **Pita Jarupunphol:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Israel, B. A., Schulz, A. J., Becker, A. B., Reyes, C. L., Parker, E. A., & Reyes, A. G. (2026). Community-based participatory research: Evolution and significant developments. *Annual Review of Public Health*, *47*, 135–157. https://doi.org/10.1146/annurev-publhealth-082224-022223

[2] Rodríguez-Pose, A., Bartalucci, F., Lozano-Gracia, N., & Dávalos, M. (2024). Overcoming left-behindedness. Moving beyond the efficiency versus equity debate in territorial development. *Regional Science Policy & Practice*, *16*(12), 100144. https://doi.org/10.1016/j.rspp.2024.100144

[3] Mortazavi, S., Eslami, M. H., Hajikhani, A., & Väätänen, J. (2021). Mapping inclusive innovation: A bibliometric study and literature review. *Journal of Business Research*, *122*, 736–750. https://doi.org/10.1016/j.jbusres.2020.07.030

[4] Hansasooksin, S. T., & Tontisirin, N. (2018). Applying the analytical hierarchy process (AHP) approach to assess an area-based innovation system in Thailand. *Nakhara: Journal of Environmental Design and Planning*, *15*, 63–76. https://doi.org/10.54028/NJ2018156376

[5] Ramani, S. V., Athreye, S., Bruder, M., & Sengupta, A. (2023). Inclusive innovation for the BoP: It's a matter of survival! *Technological Forecasting and Social Change*, *194*, 122666. https://doi.org/10.1016/j.techfore.2023.122666

[6] Taha, K., Yoo, P. D., Yeun, C., Homouz, D., & Taha, A. (2024). A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer Science Review*, *54*, 100664. https://doi.org/10.1016/j.cosrev.2024.100664

[7] Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. *Information*, *13*(2), 83. https://doi.org/10.3390/info13020083

[8] Shyrokykh, K., Girnyk, M., & Dellmuth, L. (2023). Short text classification with machine learning in the social sciences: The case of climate change on Twitter. *PLOS One*, *18*(9), e0290762. https://doi.org/10.1371/journal.pone.0290762

[9] Alangari, N., El Bachir Menai, M., Mathkour, H., & Almosallam, I. (2023). Exploring evaluation methods for interpretable machine learning: A survey. *Information*, *14*(8), 469. https://doi.org/10.3390/info14080469

[10] Mustafa, S., & Hama Saeed, M. (2025). Empowering text classification with NLP and explainable AI for enhanced interpretability. *Journal of Electrical Systems and Information Technology*, *12*(1), 81. https://doi.org/10.1186/s43067-025-00273-2

[11] Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M., & Zschech, P. (2026). Challenging the performance-interpretability trade-off: An evaluation of interpretable machine learning models. *Business & Information Systems Engineering*, *68*(1), 159–183. https://doi.org/10.1007/s12599-024-00922-2

[12] Xu, B., & Yang, G. (2025). Interpretability research of deep learning: A literature survey. *Information Fusion*, *115*, 102721. https://doi.org/10.1016/j.inffus.2024.102721

[13] Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593. https://doi.org/10.3390/electronics10050593

[14] Saarela, M., & Podgorelec, V. (2024). Recent applications of explainable AI (XAI): A systematic literature review. *Applied Sciences*, *14*(19), 8884. https://doi.org/10.3390/app14198884

[15] Rudin, C. (2022). Why black box machine learning should be avoided for high-stakes decisions, in brief. *Nature Reviews Methods Primers*, *2*(1), 81. https://doi.org/10.1038/s43586-022-00172-0

[16] Purcell, M. (2006). Urban democracy and the local trap. *Urban Studies*, *43*(11), 1921–1941. https://doi.org/10.1080/00420980600897826

[17] Trust, P., & Minghim, R. (2024). A study on text classification in the age of large language models. *Machine Learning and Knowledge Extraction*, *6*(4), 2688–2721. https://doi.org/10.3390/make6040129

[18] Kim, J., Kim, H.-U., Adamowski, J., Hatami, S., & Jeong, H. (2022). Comparative study of term-weighting schemes for environmental big data using machine learning. *Environmental Modelling & Software*, *157*, 105536. https://doi.org/10.1016/j.envsoft.2022.105536

[19] Zangari, A., Marcuzzo, M., Rizzo, M., Giudice, L., Albarelli, A., & Gasparetto, A. (2024). Hierarchical text classification and its foundations: A review of current research. *Electronics*, *13*(7), 1199. https://doi.org/10.3390/electronics13071199

[20] Naseeb, A., Zain, M., Hussain, N., Qasim, A., Ahmad, F., Sidorov, G., & Gelbukh, A. (2025). Machine learning- and deep learning-based multi-model system for hate speech detection on Facebook. *Algorithms*, *18*(6), 331. https://doi.org/10.3390/a18060331

[21] Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, *12*, 820–836. https://doi.org/10.1162/tacl_a_00675

[22] Ghanem, M., Ghaith, A. K., El-Hajj, V. G., Bhandarkar, A., de Giorgio, A., Elmi-Terander, A., & Bydon, M. (2023). Limitations in evaluating machine learning models for imbalanced binary outcome classification in spine surgery: A systematic review. *Brain Sciences*, *13*(12), 1723. https://doi.org/10.3390/brainsci13121723

[23] Hand, D. J., Christen, P., & Kirielle, N. (2021). F*: An interpretable transformation of the F-measure. *Machine Learning*, *110*(3), 451–456. https://doi.org/10.1007/s10994-021-05964-1

[24] Tarekegn, A. N., Giacobini, M., & Michalak, K. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition*, *118*, 107965. https://doi.org/10.1016/j.patcog.2021.107965

[25] Naveed, S., Stevens, G., & Robin-Kern, D. (2024). An overview of the empirical evaluation of explainable AI (XAI): A comprehensive guideline for user-centered evaluation in XAI. *Applied Sciences*, *14*(23), 11288. https://doi.org/10.3390/app142311288

[26] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms.

Neural Computation, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

[27] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*(1), 1–30. https://jmlr.org/papers/v7/demsar06a.html

How to Cite: Sudjarid, W., & Jarupunphol, P. (2026). Community-Based Innovation for Poverty Alleviation: A Text-Based Audit of Area-Based Research Funding in Thailand. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS620210062