LegitimNarrate: A dataset for analyzing legitimation mechanisms in crowdfunding narratives

Asmaa Lagrid*, Sébastien Fournier*, Bénédicte Aldebert†, Ali Ghods†, Daisy Bertrand†, Gael leboeuf†
*Lis Laboratory, †Cergam Laboratory
Aix-Marseille University, France
{asmaa.lagrid, sebastien.fournier, benedicte.aldebert, ali.ghods, daisy.bertrand, gael.leboeuf}@univ-amu.fr

Abstract

New ventures face challenges due to their liability of newness and need to gain legitimacy within the context of crowdfunding to secure vital resources for growth and survival. Previous studies have primarily assessed crowdfunding success through structured metadata or social media analytics, often neglecting detailed examinations of campaign narratives. To fill this gap, we introduce LegitimNarrate, an expert-annotated dataset specifically designed to analyze legitimation mechanisms in crowdfunding narratives. This dataset comprises 97 Kickstarter campaign descriptions segmented into 4,954 sentences, each meticulously annotated by management experts according to theoretical legitimacy frameworks. We benchmark LegitimNarrate with various contextual sentence-classification methods. This resource facilitates comprehensive research on discursive legitimacy and the role of narrative in crowdfunding contexts. The dataset is publicly available at anonymous-coder31/LegitimNarrate_dataset.

Keywords: Legitimacy, Crowdfunding, New ventures, Narratives, Dataset Annotation

1. Introduction

Entrepreneurial activity is widely recognized as a key driver of economic growth, job creation, and social innovation. Nonetheless, many new ventures fail in their early years, often due to their small size, limited resources, and insufficient market validation (Frydrych et al., 2014). Recent global entrepreneurship studies (GEM)[1] indicate that about half of all startups struggle to secure stable financing at inception, underscoring the need for alternative fundraising mechanisms.

In this context, crowdfunding has attracted growing attention for its potential to democratize access to capital by enabling ventures to engage the market, validate their business models, and attract both funding and consumer interest (Belleflamme et al., 2014; Frydrych et al., 2014). Crowdfunding platforms offer flexibility beyond traditional fundraising methods, providing opportunities to reach broader, geographically dispersed audiences (Mollick, 2014). Platforms like Kickstarter simultaneously facilitate funding and cultivate early market interest, evidenced by substantial financial contributions from millions of backers globally (Oo et al., 2023). Nevertheless, many campaigns fail to meet their funding targets due to insufficient credibility and perceived legitimacy, issues exacerbated by information asymmetry between entrepreneurs and backers (Courtney et al., 2017; Suchman, 1995).

Prior computational studies addressing crowdfunding success primarily utilize structured metadata—such as pledge amounts, number of backers, and campaign durations—or multimodal features (Tang et al., 2022; Pekar et al., 2024; Gunduz, 2024; Zhang and Lau, 2024). For instance, Zhang and Lau (2024) proposed a multimodal deep learning model that jointly integrates textual descriptions, video clips, and metadata. They found that the combination of project description, accompanying video, number of backers, and the number of previously supported projects were the most predictive features. Similarly, Tang et al. (2022) and Lee et al. (2018) employed deep attention networks to extract salient textual, visual, and audio features, demonstrating their influence on campaign outcomes. Zhou et al. (2018) explored readability and tone in project descriptions, while other studies such as Zhao et al. (2025) and Oo et al. (2023) used speech act theory to analyze how entrepreneurial language drives backer engagement. However, these works typically provide limited annotation, lack theoretical grounding in entrepreneurship, or do not make datasets publicly available—thereby restricting reproducibility and deeper investigation.

Critically, existing computational frameworks predominantly focus on linguistic structures without integrating legitimacy perspectives that are central to entrepreneurial success. Legitimacy theory emphasizes how strategic communication and narrative framing help new ventures overcome their inherent liabilities (Lounsbury and Glynn, 2001; Vaara et al., 2024). Clear legitimacy signals in campaign narratives have been shown to positively influence crowdfunding outcomes (Taeuscher et al., 2021; Chen, 2023), yet no comprehensive supervised

[1] https://www.gemconsortium.org/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1361–1371
11–16 May 2026.

dataset currently exists to analyze specific legitimation mechanisms—such as identity construction, associative references, or organizational credibility—in a structured manner.

To bridge this critical research gap, we introduce *LegitimNarrate*, the first publicly available expert-annotated dataset explicitly designed for supervised analysis of legitimation mechanisms in crowdfunding narratives. *LegitimNarrate* comprises 97 Kickstarter campaign descriptions segmented into 4,954 sentences, each annotated by management experts according to established legitimation framework. While our study does not directly predict campaign success, it systematically identifies narrative elements contributing to project credibility, significantly advancing our understanding of how legitimacy shapes crowdfunding outcomes.

The contributions of this work are as follows:

1. We introduce *LegitimNarrate*, an expertly annotated corpus with fine-grained legitimacy labels, which constitutes a foundational resource for interdisciplinary research at the intersection of computational linguistics and entrepreneurial theory;

2. We formulate two central tasks for legitimacy analysis in crowdfunding narratives: binary classification for legitimacy-signal detection and multi-class classification for the identification of specific legitimacy mechanisms;

3. We benchmark several contextual sentence-classification architectures on these tasks, thereby demonstrating the usefulness of the dataset and establishing strong baselines for future research.

2. Related Work

Several public datasets have been released to support research on crowdfunding, with most focusing on structured campaign metadata. The WebRobots Kickstarter dataset[2] contains detailed information on Kickstarter projects, including goals, funding status, and timelines. Similarly, the dataset provided by Rajat Jaiswal[3] offers metadata at the campaign level. Another dataset on Kaggle[4] includes binary labels indicating whether a campaign was successful. These datasets are useful for quantitative analysis of crowdfunding dynamics but lack annotations of narrative content. In contrast, *LegitimNarrate* provides sentence-level annotations of campaign descriptions, focusing on how entrepreneurs use narrative strategies to build legitimacy.

In related domains, some datasets target financial sentiment analysis, drawing from social media and investor communications to assess market trends, IPO activity, or company reputation (Daudert, 2022; Malo et al., 2014; Gaillat et al., 2018; Maia et al., 2018; Cortis et al.). These resources are designed to capture public sentiment and financial signals. However, *LegitimNarrate* takes a different approach. Rather than measuring financial sentiment or valuation, it examines how early-stage ventures use storytelling to gain legitimacy from potential backers. This dataset introduces a new perspective on investment-related narratives by aligning with legitimacy theory instead of traditional financial indicators, as legitimacy theory explains how new ventures gain acceptance from key audiences and reflects *a generalized perception that an entity's actions are desirable, proper, or appropriate within a socially constructed system of norms and beliefs* (Suchman, 1995). The dataset is therefore aligned with theorized mechanisms through which new ventures counter liabilities of newness and smallness.

Research distinguishes different ways organizations discursively do legitimacy—through authorizations, rationalizations, moral evaluations, and narratives (Vaara et al., 2024). Building on this tradition, entrepreneurship studies document how ventures craft claims that signal *who they are ?* (identity), *how they are organized ?* (organizational credibility), and *with whom they are associated ?* (associative ties) to mobilize resources (Überbacher, 2014; Fisher, 2020). For illustration: identity claims highlight founder's expertise and the distinctiveness of the product (e.g., *"We are a team of experienced robotics engineers"*); organizational claims often reference governance or quality routines (e.g., *"ISO-certified manufacturing partner; transparent delivery schedule"*); associative claims tie the venture to reputable actors or categories (e.g., *"Backed by a top accelerator; featured at CES"*). Empirically, clear legitimacy signals are linked to improved crowdfunding performance (Taeuscher et al., 2021; Chen, 2023; Courtney et al., 2017).

Cultural entrepreneurship emphasizes that stories and symbols are central to resource acquisition (Lounsbury and Glynn, 2001). Ventures gain attention and support by producing coherent narratives that position their offerings within recognizable categories while also highlighting distinctiveness (Navis and Glynn, 2011). Narrative devices—such as problem–solution arcs, origin stories, endorsements, and awards—help audiences make sense of novel ideas and evaluate credibility.

[2] https://webrobots.io/kickstarter-datasets

[3] https://github.com/rajatj9/Kickstarter-projects/tree/master/dataset

[4] https://www.kaggle.com/datasets/codename007/funding-successful-projects

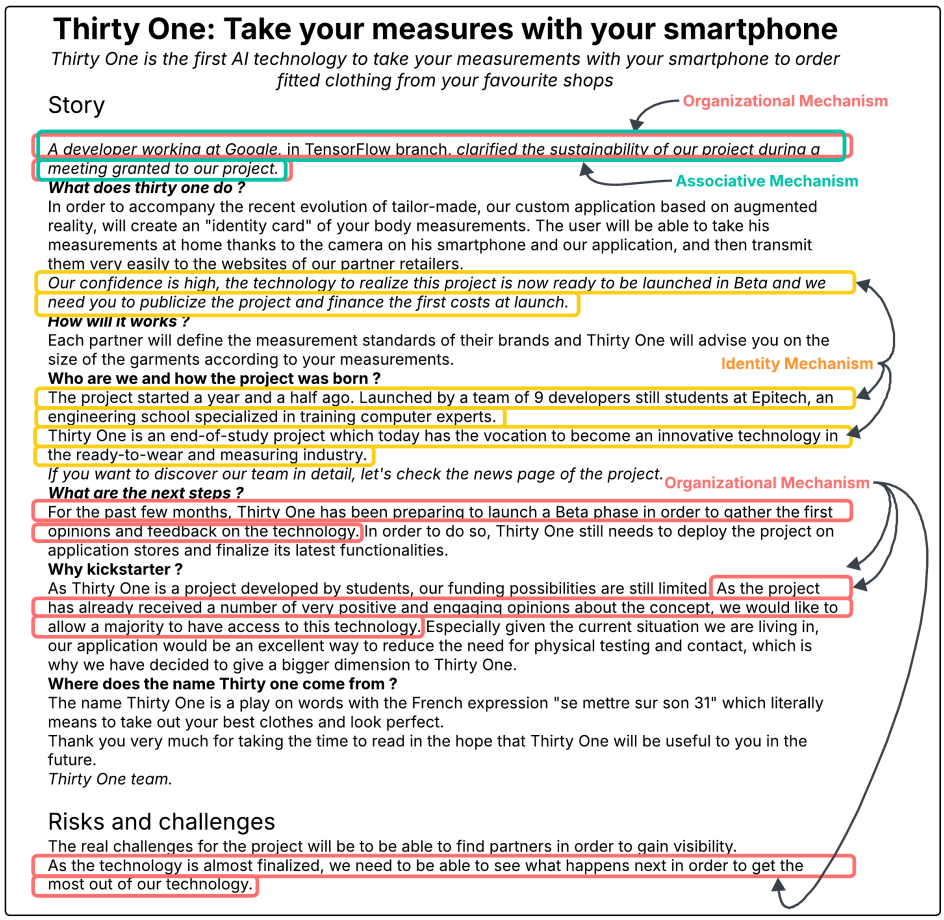

Figure 1: An example of an annotated crowdfunding project

This perspective complements legitimacy theory by specifying how sense-making and cultural frames enable acceptance. *LegitimNarrate* operationalizes this link by annotating sentence-level legitimation mechanisms that commonly appear in entrepreneurial stories (identity, organizational, associative), thus providing a practical bridge between cultural-entrepreneurship concepts and computational analysis. This design supports supervised learning for (i) detecting legitimacy cues and (ii) classifying specific mechanisms, thereby facilitating reproducible, theory-informed analysis of crowdfunding narratives and complementing prior multimodal approaches (Zhang and Lau, 2024; Tang et al., 2022).

3. Dataset Construction

This section defines the data collection, explains the mechanisms of legitimation used for the annotation, describes the annotation process, the quality assurance of annotation, and details the resulting dataset.

3.1. Corpus collection

The dataset was sourced in 2021 from the Kickstarter platform [5], focusing on technology-based crowdfunding campaigns lunched in France, and presented in English with funding goals exceeding 5,000 euros. This threshold was deliberately set to ensure the inclusion of projects demonstrating substantial commitment and detailed planning, which are typically associated with higher financial requirements.

After extracting the descriptions of projects, the text was systematically segmented into individual sentences using punctuation marks as delimiters. This sentence-level segmentation was chosen over analysis at the project-level to enable a more precise detection of legitimation mechanisms.

3.2. Category definitions

To analyze how new ventures construct legitimacy in crowdfunding narratives, we developed an annotation scheme grounded in entrepreneurial legitimacy theory. We adopt the typology introduced by

[5] https://www.kickstarter.com/?ref=nav

Fisher et al. (2017), which builds upon earlier work by Aldrich and Fiol (1994) and Delmar and Shane (2004). This framework identifies distinct mechanisms that entrepreneurs use to gain legitimacy.

Following this literature, we designed a two-tier annotation strategy. First, each sentence was annotated using a binary label to indicate whether it expresses a legitimation signal. Second, for sentences containing such a signal, we applied a multi-class label specifying the type of mechanism used.

Category 1. Legitimacy-related: Texts that fall under one of the following categories:

- **Category 1.1 - Identity mechanism:** This mechanism highlights how new ventures manage and construct their identities using cultural tools (Fisher et al., 2017) such as storytelling, impression management, and the use of arguments and analogies. These tools are vital for aligning the venture with the identity expectations of its audience, covering both explicit claims (Fisher et al., 2017) and the subtle impressions conveyed by the entrepreneur. This mechanism is crucial for establishing a resonant image that captivates potential backers. For example, Cornelissen and Clarke (2010) and Wuillaume and Janssen (2020) argue that analogies and metaphors help entrepreneurs link new ventures with familiar concepts, enhancing sensemaking. *Illustration (Figure 1):* "The project started a year and a half ago, launched by a team of nine developers still students at Epitech, an engineering school specialized in training computer experts" — this expresses *who* the founders are and their institutional affiliation.

- **Category 1.2 - Associative mechanism:** This mechanism focuses on a venture's relationships and affiliations with external entities (Fisher et al., 2017). It involves establishing ties that may include organizational links or connections with top management and influential industry figures. Entrepreneurs need to emphasize the venture's potential contributions to the community and its value to community members, often aligning with other reputable organizations to validate these contributions (Fisher et al., 2017). This association helps to bolster the venture's credibility and community standing. *Illustration (Figure 1):* "A developer working at Google (TensorFlow) clarified the feasibility of our project during a meeting" — this example primarily illustrates an associative mechanism, as the venture highlights an active relationship with a high-status external expert through direct interaction and meetings, thereby signaling proximity to a reputable actor in the field. At the same time, the expert's clarification of the project's feasibility also foreshadows the organizational mechanism discussed next, since the association is not merely symbolic but also produces concrete evaluative feedback that reinforces the venture's technical credibility.

- **Category 1.3 - Organizational mechanism:** Related to the venture's structural and performance-based legitimacy, this mechanism encompasses achieving milestones and gaining external validation (Fisher et al., 2017). Ventures must demonstrate their ability to deliver on promises, such as showcasing a prototype (Wuillaume and Janssen, 2020) or reaching specific organizational goals like product launches (Fisher et al., 2017) or obtaining certifications (Rao, 1994). The presence of qualified leadership (Fisher et al., 2017) and the implementation of a professional organizational structure (Khaire, 2010) also play critical roles in substantiating the venture's operational competence. *Illustrations (Figure 1):* "For the past few months, Thirty One has been preparing to launch a *beta* phase in order to gather initial opinions and feedback on the technology" – this highlights a concrete milestone (*beta* readiness) and delivery capability, while also reflecting operational routines (preparation and feedback collection), thus signaling execution discipline.

Category 2. Not legitimacy-related: Texts that do not fall under the three previous categories are included here.

These categories are meticulously chosen based on their prevalence in the literature and their applicability to the crowdfunding context, which often requires ventures to quickly establish a credible and engaging presence to attract funding and support.

3.3. Annotation Process

Our expert-annotated dataset was developed through a structured, multi-stage process, as following:

Stage 1. Hand annotation: Four experts, working in pairs, conducted blind annotations on each sentence to identify legitimacy-related content. Each sentence was annotated by two experts, and a third expert reviewed all sentences with disagreements to make a final classification based on majority rule among the three annotators. This initial phase yielded an imbalanced dataset with 234 "Legitimacy-related" and 1666 "Not legitimacy-related" sentences, resulting in an imbalance ratio of 7.11.

Stage 2. Data augmentation and *manual* verification: In the second stage, we mit-

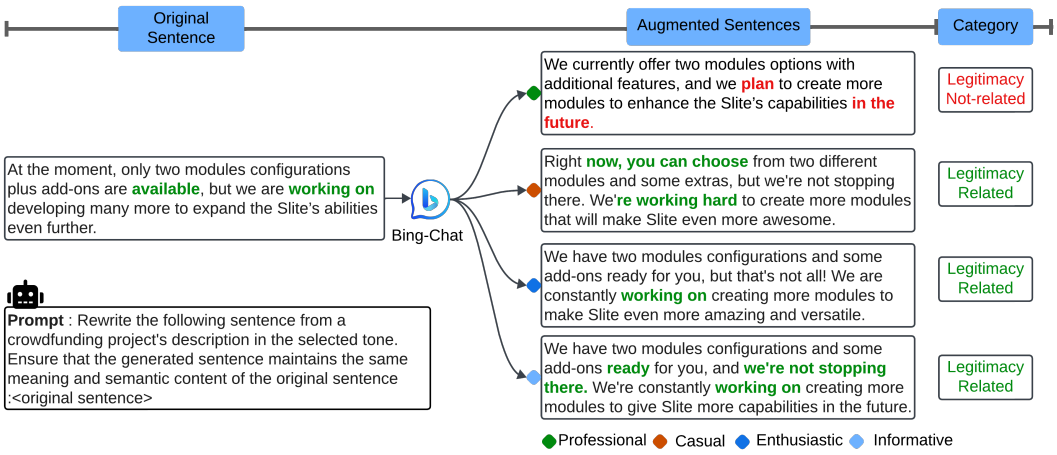

Figure 2: Data augmentation with Bing-Chat

igated class imbalance by augmenting the minority class—namely, the *legitimacy-related* sentences—via paraphrasing. We used Bing Chat (GPT-4–based) to rewrite existing sentences while preserving their original meaning. Each sentence was rendered in one of four tones: *Professional* (formal, reserved), *Casual* (conversational, relaxed), *Enthusiastic* (lively, expressive), or *Informative* (fact-focused). The generation followed a structured zero-shot prompt: *"Rewrite the following sentence from a crowdfunding project's description in the selected tone. Ensure that the generated sentence maintains the same meaning and semantic content of the original sentence: ⟨original sentence⟩"* (see Figure 2).

All augmented sentences were then manually reviewed by domain experts to ensure that both semantic content and the assigned legitimation mechanism were preserved. During this review, we observed that changes in phrasing can alter a sentence's rhetorical function. In particular, replacing operational language (e.g., "working on developing...") with vague or future-oriented expressions (e.g., "plan to create..") often removed the intended legitimation signal. As illustrated in Figure 2, such shifts can eliminate cues that are critical to organizational legitimacy. Consequently 7% of the generated sentences lost their legitimacy mechanism.

After deduplication and removal of paraphrases that lost their legitimacy signal, we retained 583 verified, annotated legitimacy-related sentences. We treat this augmented set as an auxiliary dataset and it was used only for model training in Stage 3 and is kept separate from the final released corpus. More generally, this auxiliary set may also be useful in future work for training under class imbalance, since the task naturally exhibits a long-tailed distribution.

Stage 3. Model-assisted annotation and expansion: In the final stage of our annotation process, we fine-tuned a BERT-based model on the auxiliary augmented data combined with the manually annotated data. The training results are reported in Table 1, showing that data augmentation improved the model's macro F1-score on the test set from 0.69 to 0.71. This model was then used to annotate an additional 3,054 sentences extracted from the descriptions of 61 new campaigns. Each model-predicted sentence was subsequently reviewed by an expert; in cases of disagreement, a second expert evaluated the sentence, and the final annotation was determined by majority vote among the two annotators and the model.

The resulting 3,054 sentences were merged with the original unaugmented dataset of 1,900 sentences, yielding a comprehensive collection of 4,954 sentences from 97 project descriptions, with imbalance ratio of 7.86.

Stage 4. Legitimation mechanism identification: After identifying the sentences that conveyed a legitimacy signal, we collected the mechanism labels assigned by each annotator in order to resolve cases where annotators agreed on the legitimacy-related nature of a sentence but disagreed on the specific mechanism involved. Disagreements regarding mechanism type accounted for 22% of the legitimacy-related sentences. These cases were discussed collectively to reach a final decision. Following this adjudication process, we introduced two levels of annotation: a dominant mechanism and, when relevant, a secondary mechanism. Overall, 3% of the legitimacy-related sentences were assigned a secondary mechanism.

This rigorous process ensured the accuracy and reliability of our annotated dataset and enables multiple uses—binary, multi-class, and multi-label classification—as detailed in Section 4.

Model	Validation set	Test set
Bert	0.73	0.69
Bert+aug	**0.77**	**0.71**

Table 1: Macro F1-scores for model used in Stage 3 of annotation

3.4. Annotation & Quality Assurance

The dataset was annotated by four management experts, all of whom are co-authors of this paper. Annotators A and B are professors specializing in Entrepreneurship (with A as a full professor and B as an associate professor), Annotator C is an associate professor in Finance, and Annotator D is a research engineer with expertise in Management and Entrepreneurship. Prior to annotation, Annotators C and D participated in several training sessions with Annotators A and B, who possess extensive expertise in the entrepreneurial legitimacy construct. These sessions involved reviewing examples and counterexamples for each category and subcategory in the labeling framework, ensuring a clear and consistent understanding of the annotation task.

In the initial manual annotation round, the annotations achieved a Cohen's Kappa score of 0.54 (observed agreement: 0.88; expected agreement: 0.75), indicating moderate agreement and reflecting the inherent interpretive challenges of the task. For example, the sentence "The first step to enable everyone to use the app is by translating it into as many languages as possible" was interpreted by Annotator 1 as reflecting an organizational mechanism, since it suggests concrete actions undertaken to adapt the application to a broad audience and thus signals implementation capability. In contrast, Annotator 2 considered it merely descriptive of the product's features rather than a legitimacy signal. After further review, Annotator 3 confirmed that the sentence did not express a legitimacy-related mechanism and should therefore be categorized as not legitimacy-related. Similarly, for the sentence "It is the combination of French craftsmanship, luxurious Italian full-grain calfskin, and the newest technologies in the wireless charging field", Annotator 1 identified a legitimacy signal because the sentence associates the venture with culturally valued attributes and technological sophistication. Annotator 2, however, interpreted it primarily as a storytelling or promotional element. Following adjudication, Annotator 3 confirmed that the sentence was indeed legitimacy related, as it conveys signals intended to enhance the venture's credibility and perceived value.

During the model-assisted annotation phase, the extension portion of the dataset was divided into four parts. Each part was first annotated by one expert, whose annotation was treated as the equivalent of Annotator 1 in the first annotation round. In cases where the model prediction disagreed with this first expert annotation, the sentence was then reviewed by a second expert, different from the first, who acted as Annotator 2. Thus, the "second expert" was not always the same individual, but rather one of the remaining experts involved in the annotation process. Under this protocol, the Cohen's Kappa between model predictions and first-pass human annotations (Annotator 1) reached 0.66. Importantly, in disagreement cases, 97% of the model's labels were confirmed by the second expert's review (Annotator 2), suggesting that the model's decisions were not tied to systematic alignment with a single annotator, but were instead broadly consistent with expert judgment across annotators.

All disagreements were ultimately resolved through a third round of annotation involving majority voting. For particularly challenging examples, the four annotators engaged in group discussions, highlighting the complexity of the task.

Each sentence was assigned a binary label indicating whether it was legitimacy-related or not. For those deemed legitimacy-related, annotators also specified the legitimation mechanism; agreement on mechanism labels achieved a Cohen's Kappa score of 0.83. In cases of disagreement regarding the type of legitimation mechanism, a third annotator adjudicated. When multiple categories were judged plausible, we retained a multi-label annotation, designating both a dominant and a secondary category to reflect narrative complexity.

3.5. Dataset statistics

The final dataset comprises 4,954 human-annotated sentences, with 4,395 labeled as "Not legitimacy-related" and 559 as "Legitimacy-related". Within the "Legitimacy-related" category, there are 239 sentences categorized under the "Identity Mechanism", 89 under the "Associative Mechanism", and 294 under the "Organizational Mechanism".

Notably, the dataset reveals that a single sentence can express multiple facets. For example, the sentence from the project Cakewalk 3D, "Over the past weeks, we gathered feedback from makers to food-lovers across various fablabs, incubators, and cooking schools", illustrates the Identity Mechanism by showing the venture's active engagement with a broad community to improve its offerings. This engagement reflects a commitment to inclusively and ongoing enhancement, aligning with societal values of collaboration and user-centric design. Additionally, this sentence also employs the Associative Mechanism by enhancing the venture's credibility through affiliations with recognized institutions,

Category	Count
Not Legitimacy-related	4,395
Legitimacy-related (Total)	559
Identity Mechanism	239
Organizational Mechanism	294
Associative Mechanism	89
Total	4,954

Table 2: Sentence label distribution

Length	Mean	Median	Std	Min	Max
Sentences/Project	51.07	45.00	32.98	7	168
Tokens/Sentence	19.65	17.00	12.75	2	174
Tokens/Not-related	18.98	17.00	12.914	2	166
Tokens/Identity	24.16	22.00	11.62	5	72
Tokens/Organizational	25.54	22.00	18.67	4	174
Tokens/Associative	27.19	22.00	20.99	10	174

Table 3: Descriptive statistics of sentences and projects

which helps to strengthen its legitimacy. The Table 2 shows the distribution of labels which helps in understanding the prevalence of each mechanism in crowdfunding campaign descriptions. Additionally, Table 3 provides statistics on the sentence lengths across various categories of the dataset.

4. Potential use cases

Given the information included in the *Legitim-Narrate* dataset, we envision several potential use cases for narrative analysis in the context of crowdfunding, covering both application-driven and research-driven scenarios.

4.1. Application-driven scenarios

- **Scenario.1 Legitimation mechanism detection and Scoring:** The dataset supports supervised *binary, multi-class, and multi-label classification* to identify sentences that express legitimacy signals and assign mechanism-specific labels, enabling sentence-level and document-level legitimacy scoring; it also supports *contextual sentence classification*, where predictions leverage neighboring sentences within the project description.
- **Scenario.2 Optimizing entrepreneurial communication**: By analyzing the legitimation mechanisms employed in campaign narratives, entrepreneurs can derive actionable insights to refine their communication strategies. This enables them to better align their messages with the expectations and norms of potential investors and customers, potentially increasing their chances of funding success.
- **Scenario.3 Recommendation system for crowdfunding projects:** The dataset can be used to build an investment recommender that ranks and suggests campaigns based on their estimated legitimacy score.

4.2. Research-driven scenarios

- **Scenario.4 Impact analysis on campaign success:** Researchers can integrate our dataset with other datasets that contain comprehensive campaign metadata to assess the influence of legitimacy scores on the overall success of crowdfunding campaigns.
- **Scenario.5 Ai-generated sentences detection :** During the data augmentation and manual verification phase (Section 3.3), we constructed a small auxiliary dataset consisting of paraphrases produced by Bing-Chat (GPT-4) for minority-class (legitimacy-related) sentences. This auxiliary material—not included in the released corpus—enables a binary classification task contrasting human-authored (original) sentences with LLM-generated (paraphrased) ones.
- **Scenario.6 Learning with class imbalance:** The corpus displays a genuine long-tailed distribution, with imbalance ratios of 7.86 for binary classification and 46.3 for multi-label classification. This characteristic makes it particularly suitable for investigating training and evaluation strategies under class imbalance.

As the main use case considered during the construction of this dataset, we provide experiments on legitimation mechanism detection and legitimacy scoring in the following section.

5. Experimental evaluation

To conduct our experiments, we partitioned the corpus at the project level into three splits: a test set (10 projects), a validation set (6 projects), and the remaining projects for training. Selection was stratified by the prevalence of multi-label annotations, with projects containing many multi-label sentences deliberately assigned to the test set to increase difficulty. The label distribution across splits is shown in Figure 3.

5.1. Scenario 1: Illustrative Experiment

We conduct two sets of experiments on Scenario 1 where we evaluate the effectiveness of contextual sentence classification models on our dataset.

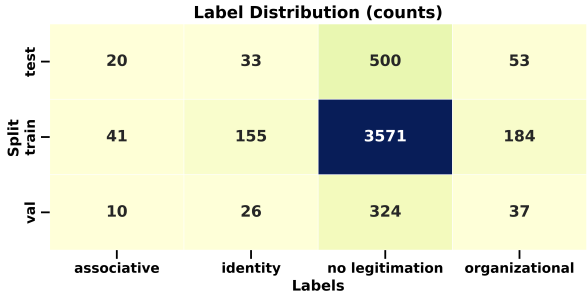

Figure 3: Label distribution among different data splits

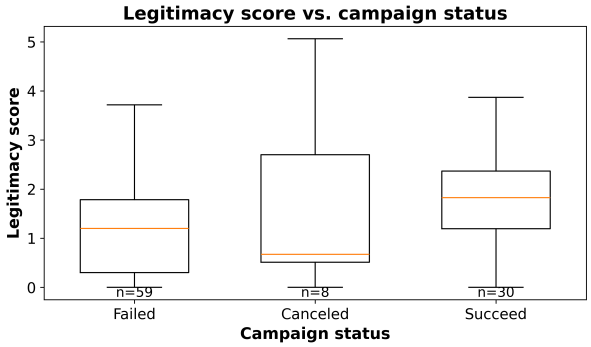

Figure 4: Campaign status vs legitimacy score

The first task is a binary classification, where sentences are categorized as either legitimacy-related or not legitimacy-related. The second task is a more fine-grained multi-class classification, where each sentence is classified into one of several classes: not-related or one of the specific legitimation mechanisms. In this setting, we consider only the dominant mechanism assigned to each legitimacy-related sentence, without modeling the task as multi-label classification. Specifically, we compare two architectures: (1) a global-context model encodes each sentence with BERT-Large, then fuses these representations with a project-level context produced by ModernBert-Large (Warner et al., 2024); and (2) a local context model based on Longformer-large (Beltagy et al., 2020) using a sliding window of 10 sentences. In addition, we assess the zero-shot classification capabilities of LLama3-8b (AI@Meta, 2024), and Mistral 8x22B (Jiang et al., 2024), representing recent large language models, and we include ModerBert-large as baseline model without a contextual classification.

Results are reported in Table 4. The Longformer-based model outperforms all other approaches, achieving a F1-score(Macro) of 0.72 on the binary classification task and 0.53 on the multi-class task. These results underscore the importance of modeling local sentence context when detecting subtle rhetorical mechanisms. Notably, this task remains challenging due to the contextual dependency of legitimacy cues, as reflected in inter-annotator disagreement during the labeling phase.

Model	Binary	Multi-class
ModernBert	0.67	0.47
Longformer	**0.72**	**0.53**
ModernBert-Bert	0.70	0.47
Mixtral8x22B	0.62	0.39
LLaMA3-8B	0.53	0.32

Table 4: Macro F1-scores for each model (Best per task in bold)

5.2. Scenario 4: Illustrative Experiment

Beyond classification, the dataset records the funding status of each campaign, which enables an exploratory analysis of the relationship between discursive legitimacy and performance. We define a simple *legitimacy score* at the project level. For project i, let M_i denote the number of sentences annotated as expressing at least one legitimation mechanism (identity, organizational, or associative), and let S_i be the total number of sentences in the project description. The score is:

$$L_i = \frac{M_i}{\log(1+S_i)}.$$

This length-normalized ratio down-weights very long descriptions without overly penalizing them.

Using this score, preliminary results (see Figure 4) indicate that successful campaigns tend to exhibit a higher average L_i than failed ones, suggesting that a stronger presence of legitimation mechanisms may be associated with more favorable crowdfunding outcomes. Interestingly, the highest score was observed for a canceled project, which suggests that campaign cancellation may depend on factors beyond narrative legitimation, such as media coverage or campaign duration. This measure remains intentionally simple, and future work could refine it by introducing mechanism-specific weights (e.g., $L_i^{(w)} = \frac{\sum_k w_k M_{ik}}{\log(1+S_i)}$ with $k \in \{\text{identity, organizational, associative}\}$) or project-type-dependent priors.

6. Discussion

Overall performance remains modest, which is expected given the subtlety of legitimacy signalling at the sentence level. Signals often hinge on fine-grained wording and local context, while the semantic boundaries between mechanisms are narrow. Sentence-level inputs also ignore project-level cues (e.g., preceding sentences naming a partner or a certification) that are frequently required

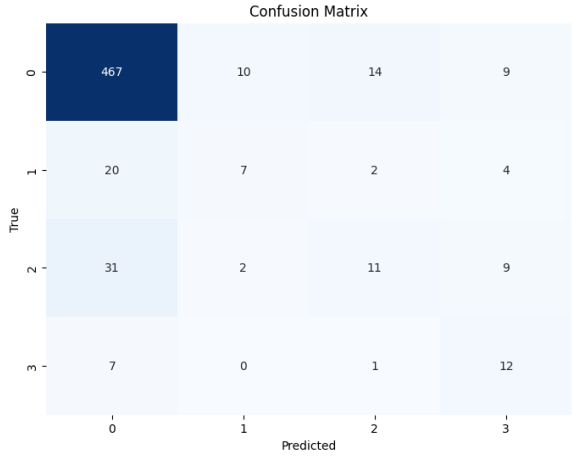

Figure 5: Confusion matrix of the baseline (ModernBert)

to disambiguate a claim. In addition, class imbalance—where legitimacy mechanisms are comparatively rare—tends to induce over-prediction of the majority class ("not related-legitimacy") and under-detection of minority classes.

The confusion matrix of the baseline illustrated in Figure 5 indicates that most errors occur between the *not related-legitimacy* class and *legitimacy-related* classes, with fewer but notable confusions among the mechanisms themselves. A recurring false positive arises when the model treats qualitative product descriptions as evidence; for example, "Besides being a very nice product, Zimpure is now very quiet" is descriptive, but it does not provide a verifiable metric, a recognized third-party association, or an organizational proof, and should therefore be labeled as *not related-legitimacy*. Conversely, typical false negatives include identity-oriented sense-making that the model dismisses as mere storytelling, such as "One day, while walking along the Seine in Paris at sunset, we got the idea to create a clock with the binary system where light indicate digits, whose shape would be a Parisian building where we could imagine people living as lights turn on and off", which frames vision and purpose and thus constitutes an identity signal. In the multi-class classification task, some sentences identified as false positives or false negatives for a given mechanism may in fact express more than one legitimation mechanism. In such cases, the model may fail to predict the dominant mechanism, even though the sentence contains valid cues for another category. This difficulty arises because a single sentence can simultaneously convey multiple legitimacy signals. As illustrated in Figure 1, the first sentence reflects both organizational and associative mechanisms, but the model may capture only one of them while missing the dominant label

These observations suggest three modeling directions. First, a multi-label formulation can better capture co-occurring mechanisms within a sentence. Second, more sophisticated context-aware architectures—such as sequence tagging with cross-sentence attention—can incorporate nearby cues (e.g., the sentence that names a partner or cites a standard). Third, imbalance-aware learning, including class-aware sampling, re-weighting, mixup/remix for minority classes, and per-class threshold tuning with calibration, can mitigate skew without distorting sentence wording. Complementary few-shot approaches with concise definitions and counterexamples may also help large language models discriminate borderline cases.

Several limitations point to directions for future work. Expanding coverage of minority classes and incorporating domain-specific lexicons (e.g., standards, grants, and institutions) may improve precision and recall for legitimacy-related categories. Moreover, evaluating document-level models that jointly capture local and global context could help reduce a substantial portion of the observed false positives and false negatives.

7. Conclusion

We introduced *LegitimNarrate*, the first expert-annotated dataset dedicated to analyzing how new ventures construct legitimacy through narrative in crowdfunding contexts. The dataset was carefully built to capture subtle linguistic signals of legitimacy, with particular attention to word choice and contextual nuance. To address the natural class imbalance—due to the rarity of legitimacy-related sentences—we applied expert-verified data augmentation.

LegitimNarrate enables a range of research applications, including legitimation mechanism classification, crowdfunding success prediction, and contextual sentence modeling. Its structure also supports critical NLP challenges such as class imbalance learning and few-shot learning, given the scarcity of positive samples. This dataset offers a foundation for advancing computational methods that require semantic sensitivity, rhetorical awareness, and discourse-level understanding.

Ethics Statement

This study uses publicly available Kickstarter campaign descriptions. Only content voluntarily disclosed by project creators was collected, and no private or sensitive personal data were used. No direct interaction with human participants was involved.

The annotations were produced by domain experts who are also co-authors of the paper. Al-

though this ensured alignment with the theoretical framework, it may introduce interpretive bias. To reduce this risk, annotation was performed independently, disagreements were reviewed through adjudication, and inter-annotator agreement was reported.

8. Bibliographical References

AI@Meta. 2024. Llama 3 model card.

Howard E Aldrich and C Marlene Fiol. 1994. Fools rush in? the institutional context of industry creation. *Academy of management review*, 19(4):645–670.

Paul Belleflamme, Thomas Lambert, and Armin Schwienbacher. 2014. Crowdfunding: Tapping the right crowd. *Journal of business venturing*, 29(5):585–609.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Wendy D Chen. 2023. Crowdfunding: Different types of legitimacy. *Small Business Economics*, 60(1):245–263.

Joep P Cornelissen and Jean S Clarke. 2010. Imagining and rationalizing opportunities: Inductive reasoning and the creation and justification of new ventures. *Academy of Management Review*, 35(4):539–557.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 519–535.

Christopher Courtney, Supradeep Dutta, and Yong Li. 2017. Resolving information asymmetry: Signaling, endorsement, and crowdfunding success. *Entrepreneurship Theory and Practice*, 41(2):265–290.

Tobias Daudert. 2022. A multi-source entity-level sentiment corpus for the financial domain: the finlin corpus. *Language Resources and Evaluation*, 56(1):333–356.

Frédéric Delmar and Scott Shane. 2004. Legitimating first: Organizing activities and the survival of new ventures. *Journal of business venturing*, 19(3):385–410.

Greg Fisher. 2020. The complexities of new venture legitimacy. *Organization Theory*, 1(2):2631787720913881.

Greg Fisher, Donald F Kuratko, James M Bloodgood, and Jeffrey S Hornsby. 2017. Legitimate to whom? the challenge of audience diversity and new venture legitimacy. *Journal of Business Venturing*, 32(1):52–71.

Denis Frydrych, Adam J Bock, Tony Kinder, and Benjamin Koeck. 2014. Exploring entrepreneurial legitimacy in reward-based crowdfunding. *Venture capital*, 16(3):247–269.

Thomas Gaillat, Manel Zarrouk, André Freitas, and Brian Davis. 2018. The ssix corpora: Three gold standard corpora for sentiment analysis in english, spanish and german financial microblogs. In *LREC: Language Resources and Evaluation Conference*, pages 2671–2675. European Languages Resources Association (ELRA).

Hakan Gunduz. 2024. Comparative analysis of bert and fasttext representations on crowdfunding campaign success prediction. *PeerJ Computer Science*, 10:e2316.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Mukti Khaire. 2010. Young and no money? never mind: The material impact of social resources on new venture growth. *Organization Science*, 21(1):168–185.

SeungHun Lee, KangHee Lee, and Hyun-chul Kim. 2018. Content-based success prediction of crowdfunding campaigns: A deep learning approach. In *Companion of the 2018 ACM conference on computer supported cooperative work and social computing*, pages 193–196.

Michael Lounsbury and Mary Ann Glynn. 2001. Cultural entrepreneurship: Stories, legitimacy, and the acquisition of resources. *Strategic management journal*, 22(6-7):545–564.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in

economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Ethan Mollick. 2014. The dynamics of crowdfunding: An exploratory study. *Journal of business venturing*, 29(1):1–16.

Chad Navis and Mary Ann Glynn. 2011. Legitimate distinctiveness and the entrepreneurial identity: Influence on investor judgments of new venture plausibility. *Academy of Management Review*, 36(3):479–499.

Pyayt P Oo, Lin Jiang, Arvin Sahaym, Annaleena Parhankangas, and Richard Chan. 2023. Actions in words: How entrepreneurs use diversified and changing speech acts to achieve funding success. *Journal of Business Venturing*, 38(2):106289.

Viktor Pekar, Marina Candi, Ahmad Beltagui, Nikolaos Stylos, and Wei Liu. 2024. Explainable text-based features in predictive models of crowdfunding campaigns. *Annals of Operations Research*, pages 1–31.

Hayagreeva Rao. 1994. The social construction of reputation: Certification contests, legitimation, and the survival of organizations in the american automobile industry: 1895–1912. *Strategic management journal*, 15(S1):29–44.

Mark C Suchman. 1995. Managing legitimacy: Strategic and institutional approaches. *Academy of management review*, 20(3):571–610.

Karl Taeuscher, Ricarda Bouncken, and Robin Pesch. 2021. Gaining legitimacy by being different: Optimal distinctiveness in crowdfunding platforms. *Academy of Management Journal*, 64(1):149–179.

Zhe Tang, Yi Yang, Wen Li, Defu Lian, and Lixin Duan. 2022. Deep cross-attention network for crowdfunding success prediction. *IEEE Transactions on Multimedia*, 25:1306–1319.

Florian Überbacher. 2014. Legitimation of new ventures: A review and research programme. *Journal of management studies*, 51(4):667–698.

Eero Vaara, Ana M Aranda, and Helen Etchanchu. 2024. Discursive legitimation: An integrative theoretical framework and agenda for future research. *Journal of Management*, page 01492063241230511.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Amélie Wuillaume and Frank Janssen. 2020. Legitimacy in crowdfunding: Some surprising patterns. *International Review of Entrepreneurship*, 18(4).

Zijian Zhang and Raymond YK Lau. 2024. Exploiting multimodal features and deep learning for predicting crowdfunding successes. In *2024 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–6. IEEE.

Yiming Zhao, Chaoqun Deng, and Qiang Gao. 2025. Fine-grained feature fusion framework for online crowdfunding success prediction.

Mi Zhou, Baozhou Lu, Weiguo Fan, and G Alan Wang. 2018. Project description and crowdfunding success: an exploratory study. *Information Systems Frontiers*, 20:259–274.

9. Language Resource References

The dataset is publicly available on GitHub at anonymous-coder31/LegitimNarrate_dataset.