algorithms

MDPI

Article

A System Design Perspective for Business Growth in a Crowdsourced Data Labeling Practice

Vahid Hajipour [1,2], Sajjad Jalali [2], Francisco Javier Santos-Arteaga [3,*], Samira Vazifeh Noshafagh [4] and Debora Di Caprio [5]

[1] Department of Industrial Engineering, West Tehran Branch, Islamic Azad University, Tehran 1468763785, Iran; vhajipour@iau.ac.ir
[2] Research Center, FANAP Co., Tehran 1657245030, Iran; s.jalali@fanap.ir
[3] Department of Financial and Actuarial Economics & Statistics, Universidad Complutense de Madrid, 28223 Madrid, Spain
[4] Department of Industrial Engineering, Doctoral Programme in Materials, Mechatronics and Systems Engineering, University of Trento, 38123 Trento, Italy; s.vazifehnoshafagh@unitn.it
[5] Department of Economics and Management, University of Trento, 38122 Trento, Italy; debora.dicaprio@unitn.it
* Correspondence: fransant@ucm.es

Abstract: Data labeling systems are designed to facilitate the training and validation of machine learning algorithms under the umbrella of crowdsourcing practices. The current paper presents a novel approach for designing a customized data labeling system, emphasizing two key aspects: an innovative payment mechanism for users and an efficient configuration of output results. The main problem addressed is the labeling of datasets where golden items are utilized to verify user performance and assure the quality of the annotated outputs. Our proposed payment mechanism is enhanced through a modified skip-based golden-oriented function that balances user penalties and prevents spam activities. Additionally, we introduce a comprehensive reporting framework to measure aggregated results and accuracy levels, ensuring the reliability of the labeling output. Our findings indicate that the proposed solutions are pivotal in incentivizing user participation, thereby reinforcing the applicability and profitability of newly launched labeling systems.

Keywords: data labeling; annotation; crowdsourcing; payment mechanism; machine learning

check for updates

Citation: Hajipour, V.; Jalali, S.; Santos-Arteaga, F.J.; Vazifeh Noshafagh, S.; Di Caprio, D. A System Design Perspective for Business Growth in a Crowdsourced Data Labeling Practice. *Algorithms* **2024**, *17*, 357. https://doi.org/10.3390/a17080357

Academic Editor: Frank Werner

Received: 9 July 2024
Revised: 9 August 2024
Accepted: 11 August 2024
Published: 15 August 2024

1. Introduction

The provision of labeled datasets in the most efficient possible way is of paramount importance to enhance the categorization capacities of machine learning methods, particularly supervised learning [1,2]. Data labeling systems and applications are enabling tools to steer annotation practices through the cooperation of a large number of volunteers conducting a set of minor tasks. Herein, the volunteers are called individuals, contributors, users, crowds, and/or workers, while the microtasks include common labeling activities such as identifying and marking specific facets of images. For instance, as illustrated in the literature, user experience is essential to designing efficient and adaptable interactive industrial Internet of Things systems [3]. Using the wisdom and cognitive ability of an undefined network of individuals to respond to an open call is also denoted as crowdsourcing [4–6]. Generally, crowdsourced data labeling refers to the process of assigning tags to raw data (such as images or text) by leveraging the contributions of a large group of people, often through online platforms.

The current paper explores a data labeling system that facilitates the crowdsourcing practice of annotating large-scale datasets from the perspectives of payment mechanism design and reporting. Together with the established aggregating metrics, the modifications adopted in the payment mechanism provide a basis for the system's development and usability to handle a vast quantity of data. The proposed techniques, recommendations, and findings are the result of implementing a real data labeling system denoted POBEL developed by a large technological solution provider called FANAP Co.

In this paper, we address two critical components in the design of an efficient data labeling system: an innovative payment mechanism and a robust configuration of output results. The payment mechanism is designed to incentivize user participation and maintain the integrity of the labeling process by incorporating a skip-based golden-oriented function. This function not only balances user penalties but also mitigates spam activities. On the other hand, the configuration of output results is managed through a comprehensive reporting framework. This framework measures the aggregated results and accuracy levels, ensuring the reliability of the annotated outputs. By focusing on these two aspects, our approach aims to enhance the overall efficiency and effectiveness of crowdsourced data labeling systems.

In a nutshell, labeling is the act of detecting particular characteristics of raw data and associating them with factual points to address their correct functionality. A typical example is the selection of correct images that have a certain feature (e.g., being an animal) among a set of given data. Pinpointing the sense of sentences constitutes another feature of the labeling activity in which positive, negative, or neutral tones are detected. Procuring labeled clean data about any concept is vital for the training and pattern discovery qualities of machine learning algorithms. At the same time, it is infeasible to assign the responsibility of labeling a high volume of data to a single or few individuals [7]. Thus, data labeling crowdsourcing-based applications are a workable solution designed to provide a platform for feeding the data required by the underlying algorithms [8]. One of the prevalent topics is how to motivate the crowd to participate in the labeling practice using financial incentives. The crowd should be paid based on their performance issuing high-quality labels, which requires a systematic payment mechanism. Such a mechanism is the central element of any data labeling system and thereby it conditions the behavior of other elements like the distribution strategy of items and the aggregation of the results' configuration.

A well-known research stream related to the management of the crowd-workers' payment mechanism is the utilization of golden items within the pool of data that is going to be annotated. Golden items are a pre-specified batch of data whose correct labels are known by the systems' admin but unrecognizable to ordinary users. By mixing golden items into the pool of data, the performance of the users can be measured by comparing the quality of their answers to the golden items [9]. The current study applies the golden approach for controlling the payment mechanism of the proposed data labeling system. In this regard, the closest study to ours relates to the skip-based approach of [10]. The authors proved that their approach was the most reliable one for satisfying the no-free-lunch axiom. This axiom hedges against paying more credit than the lowest possible one to the workers that assign incorrect labels to the golden data.

Given a number of golden items within the main pool of data, ref. [10] designed a multiplicative credit function to set the workers' payments between the pre-defined minimum and maximum thresholds. In particular, the scores of users increased exponentially based on their correct responses to the golden data. The rate of increase followed the inverse amount of the significance level. Since the significance level was between zero and one, the rate easily became greater than one. On the other hand, users' credit plummeted to the minimum threshold when assigning one wrong label to the golden items. Users could also skip the items and keep their score intact when doubtful about the correct labels of the items. This paper customizes the function proposed by [10], providing practical solutions to resolve the following concerns of their study—each of them leading to a research question that will be addressed in this study:

- *First Concern* (*C1*): Shifting the credit of users to the lowest level due to the submission of a single error seems an overly strict rule. That is, an individual who has submitted 99 correct labels to the golden items and now reserves USD X in his/her wallet can lose the whole credit by submitting one wrong label. Under such a rigorous condition, the workers of a new crowdsourcing business may consider the scoring method unfair and halt their contributions.

- o *Research Question 1* (*RQ1*): How can the credit function of [10] be modified to reasonably alleviate the rigorous condition applied to the provision of incorrect answers?
- • *Second Concern* (*C2*): Ref. [10] did not provide information about whether all golden items are supposed to be purely positive in the True/False label type. Consider an image dataset of celebrities in which individuals are asked about the conformity of a specific photo with a given celebrity's name. What would be the possible consequence(s) if all golden items were assigned True-type labels (e.g., the correct label for all the golden items was True)? It would then be possible for a spammer to assign the True label to all data and collect the whole credit since the penalty is only activated when False-type labels are assigned to golden items. In this way, the spammer would not be penalized when submitting True-type labels to non-golden data that do not belong to a given celebrity.
- o *Research Question 2* (*RQ2*): What changes related to single-type golden data must be made to avoid the cheating action of spammers?
- • *Third Concern* (*C3*): Consider now the settings of the function's parameters. The proposed credit function could become inappropriate over large-scale datasets if workable operations are not adopted to tune the underlying parameters. By neglecting parameter tuning, a user who contributes by labeling a small percentage of the entire large-scale dataset would obtain a negligible credit of less than one unit. Since the formula obeys an exponential distribution, the growing trend of the credit only becomes tangible after a significant number of dataset items have been labeled. Therefore, the user may doubt the trustfulness of the proposed data labeling system.
- o *Research Question 3* (*RQ3*): How can the credit function parameters of [10] be tuned?
- • *Fourth Concern* (*C4*): The way golden questions are distributed into the pool of ordinary data may be influential on the viewpoints of the users. For instance, suppose that 10 golden items out of 100 dataset items are shown in a row to the users. Then, a user may think that the credit function does not work properly as his or her score stops increasing when labeling the remaining 90 items. This constitutes a potential drawback derived from the distribution of golden data.
- o *Research Question 4* (*RQ4*): What kind of practical yet easily implementable distribution mechanism(s) can be employed to enhance the efficacy of the associated credit function?
- • *Fifth concern* (*C5*): The implementation of the aforementioned framework within a real data labeling system requires a specific reporting format to configure the outputs. In this regard, the corresponding payment and distribution mechanisms must deliver clean annotated data to the customer. These features have rarely been studied and need to be discussed.
- o *Research Question 5* (*RQ5*): What metrics should be considered in the reporting framework of the proposed data labeling system?

Addressing and answering the above research questions constitutes the main contribution of the current manuscript relative to previous studies. In particular, the development stage of the proposed data labeling system has been specifically designed to address the above concerns. The solutions applied resort to best practices together with heuristic and workable approaches.

The rest of the paper proceeds as follows: Section 2 reviews the literature; Section 3 explores the working structure of the proposed data labeling system and discusses the payment mechanism; Section 4 presents different sensitivity analyses and Section 5 configures a reporting template for aggregating the outputs; and Section 6 concludes and provides future research recommendations.

2. Literature Review

This section reviews the recent applications of data labeling systems, different types of payment mechanisms, as well as prospective quality control and aggregating metrics.

Research involving the design of crowdsourcing-based applications to handle the various requirements pertaining to data labeling problems has consistently grown since 2006 [11,12]. Previous studies have shed light on the system-based implementation of data labeling, including speech recognition, environmental assessment, text scan, image detection, and sentiment analysis. For instance, ref. [13] designed a web-based crowdsourcing application to obtain a large-scale speech emotion recognition dataset for easing the learning of the speaker-adaptive systems. The application allowed users to select a specific emotion, e.g., fear, per random phrase and record their voice to convey the same sense. Users also had the chance to preview the recorded voice and modify it before submission. Despite considering a convolutional neural network for transfer learning, the authors did not describe how users' annotations were validated in the proposed crowdsourcing application.

Ref. [14] used crowdsourcing to assess the post-disaster damage level of constructions by involving citizens. The microtask was to complete a questionnaire composed of a set of simple items. The predefined decision rules together with the answers obtained were used to assess the degree of raw damage. The final degree of damage was estimated through statistical inference and reported to the crisis management office to adopt the required actions and dispatch rescue forces. Ref. [15] developed a crowdsourcing web application to scan the key terms of scientific papers. The application eased the procedure of finding papers related to keywords, ranking the papers retrieved based on their impact factor while screening, annotating, and classifying the text. The data labeled were aggregated through the majority voting approach. The authors enhanced their previous design by improving the usability and efficiency of the application [16].

Recent developments focus on the strategies designed to achieve and maintain a critical mass of motivated users [17], which has led to the introduction of motivational tactics that borrow their main qualities from games, a process known as gamification [18].

There is also a rising number of studies associated with the analysis of the credit function and payment mechanism in terms of the users' performance. Ref. [19] introduced two approaches, namely, majority decision and control group, for evaluating the work submitted by users. In the majority decision approach, all the workers involved were paid, and the aggregation was made based on the most frequent response among the annotations submitted. The control group approach implemented a more rigorous method by delegating a task to a specific worker and subsequently double checking the results submitted by a group of users. If the majority of the control group confirmed the result of the initial user, he/she received the bonus. This method became more applicable when the task of the individuals within the group was cheaper and easier than that of the initial user. For instance, the initial user was supposed to write an abstract about a particular topic while the control group's users assessed the quality of the work submitted by scanning the text. The analysis performed by the authors showed that both approaches provide a significant level of confidence for detecting untrustworthy annotators. However, the majority decision worked better with low-price tasks, whereas the control group outperformed its counterpart in the case of high-price ones.

Ref. [20] defined a Nash equilibrium within the incentive mechanism of the crowdsourcing setting in order to minimize total payment. Ref. [21] suggested a dynamic distribution of the questions to minimize the active labeling duration of a spammer or careless contributors by discovering correlated performance patterns. Their results illustrated the superiority of the dynamic approach relative to the static one in terms of rework rate reduction. Ref. [22] segmented a well-defined crowdsourcing quality control taxonomy into its model, assessment, and assurance components. The authors showed that ground truth data, inclusive of golden or control questions, could fully measure the performance of users. Figuring out the malicious behavior of contributors in terms of the responses submitted to online surveys was also the research topic of [23]. These authors developed an approach to evaluate the maliciousness of the contributors and grouped the spammers into five categories ranging from ineligible workers to smart deceivers. Ref. [24] extracted two strategies from previous studies to enhance crowdsourced labeled data on the task design and after the data collection stages. In the former stage, a real-time feedback system together with a shared workflow between workers and requesters, periodical checkpoints, and a golden-led payment mechanism were all utilized to increase the quality of outputs [10,25,26]. In the latter stage, trust models together with the imposition of replication rules were employed to sift through spamming activities.

In the current study, we incorporate quality control to the task design by introducing practical solutions into the golden-led payment mechanism of [10]. The introduction of quality control following data collection fosters the application of our approach to real crowdsourcing data labeling systems. In particular, we impose a replication constraint and define reliability metrics across different aggregating report scenarios. As illustrated in Table 1, the contribution of the current paper to the literature consists in a simultaneous application of practical data labeling-based solutions. This is carried out by tuning the payment function, considering dual-type golden data, incorporating data distribution strategies, and configuring final reports into the data labeling system design practice.

Table 1. Literature highlights.

References	Orientation	Involving Business Mechanism	Tuning Payment Function	Quality Control Measures	Data Distribution	Configuring Systems' Output
[19]	Payment mechanism	×	×	Majority voting and control group	×	×
[25]	Quality enhancement	×	×	Feedback system	×	✔
[26]	Quality enhancement	×	×	Collaborative labeling	×	×
[20]	Payment mechanism	×	×	Game-theoretic tool	×	×
[23]	Quality enhancement	×	×	Spam detection	×	×
[10]	Payment mechanism	×	×	Single-type golden data	×	×
[24]	Quality enhancement	×	×	In-between and after labeling	×	×
[21]	Quality enhancement	×	×	Dynamic data distribution	✔	✔
[22]	Quality enhancement	×	×	Framing structure	×	✔
[15]	Application design	✔	×	Majority voting	×	×
[16]	Application design	✔	×	Majority voting	×	×
[9]	Payment mechanism	×	×	Single-type golden data	×	×
[13]	Application design	✔	×	-	×	×
[14]	Application design	✔	×	-	×	✔
Present study	Application design	✔	✔	Dual-type golden data	✔	✔

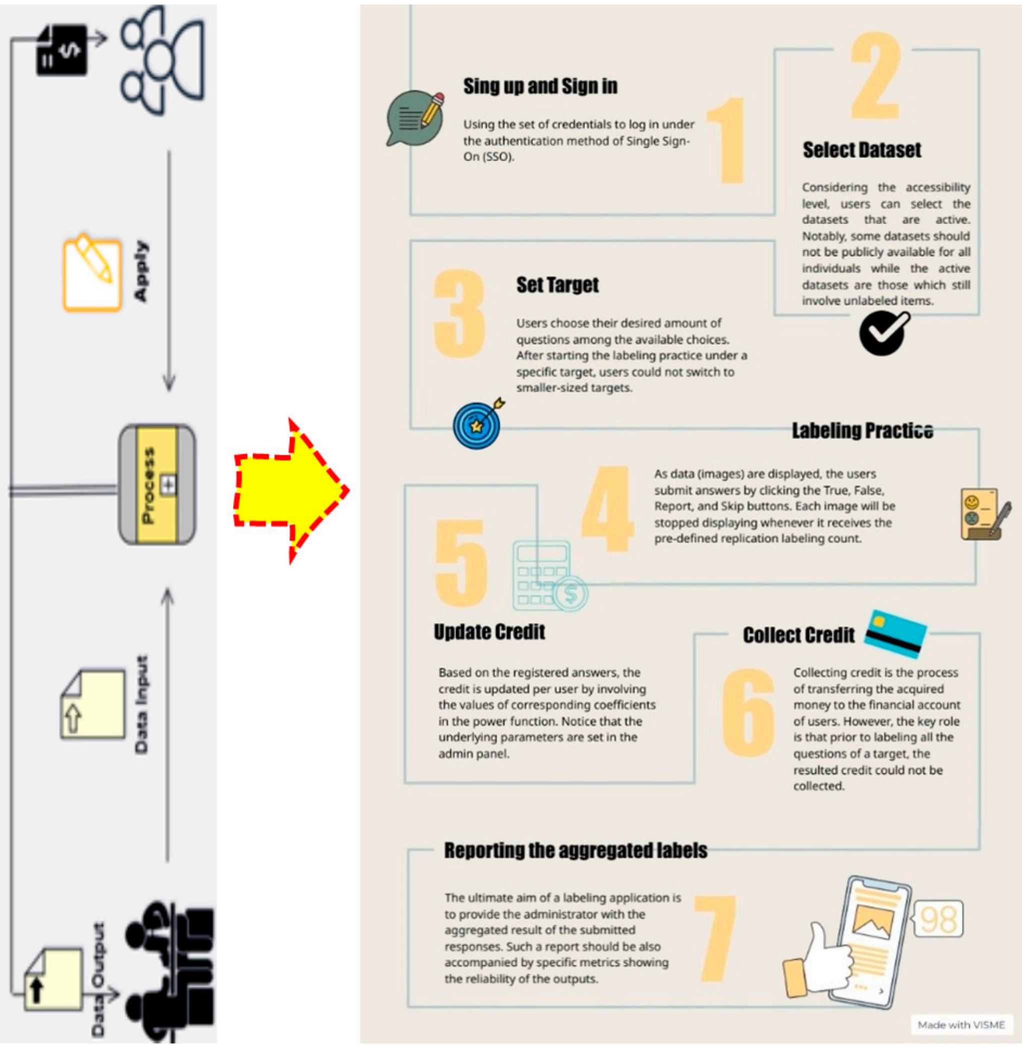
In addition to its current implementation, a variety of information-retrieval scenarios facing reliability frictions arise as potential business applications of POBEL. For instance, its features are particularly relevant in the initial development stages of firms, which require processing large amounts of user data and dealing with potential drawbacks regarding the quality of the information collected [27,28]. A similar business application would follow from the collection of reliable data from firm employees and managers to be processed through enterprise resource planning systems [29,30].

3. Payment Mechanism

This section studies the features, challenges, and solutions defined to set out an applied credit function when designing a real data labeling system. Prior to describing the mathematical formula, we analyze the functionality of the payment mechanism within the working process of the proposed system. We then describe how the basic version of the credit function evolves into a new and workable one while responding to questions *RQ1* to *RQ4*.

3.1. Position of the Payment Mechanism in the Working Process

Figure 1 describes the general process of the labeling system, emphasizing the position of its credit function. Note that all the settings described are adjustable in the admin panel. Like any other system, the labeling practice starts by receiving the input data from customers and involving the crowd workers. The system must reward workers fairly and deliver accurately labeled data to customers. The design proposed involves several consecutive steps. Users enter the system through an authentication method and select their desirable datasets from among those available. They also determine their contribution level prior to starting, a procedure known as target setting. Users can then start labeling, i.e., submitting the appropriate answer to each item. Data are displayed until the number of labels assigned to each item reaches the pre-defined replication count.



Using the set of credentials to log in under the authentication method of Single Sign-On (SSO).

Considering the accessibility level, users can select the datasets that are active. Notably, some datasets should not be publicly available for all individuals while the active datasets are those which still involve unlabeled items.

Users choose their desired amount of questions among the available choices. After starting the labeling practice under a specific target, users could not switch to smaller-sized targets.

As data (images) are displayed, the users submit answers by clicking the True, False, Report, and Skip buttons. Each image will be stopped displaying whenever it receives the pre-defined replication labeling count.

Based on the registered answers, the credit is updated per user by involving the values of corresponding coefficients in the power function. Notice that the underlying parameters are set in the admin panel.

Collecting credit is the process of transferring the acquired money to the financial account of users. However, the key role is that prior to labeling all the questions of a target, the resulted credit could not be collected.

The ultimate aim of a labeling application is to provide the administrator with the aggregated result of the submitted responses. Such a report should be also accompanied by specific metrics showing the reliability of the outputs.

Figure 1. General process of a data labeling system.

As users submit their answers, credit is updated by inputting the correct and/or incorrect responses to golden items into the proposed function. When a user completes labeling a target, he/she is allowed to collect the corresponding credit by transferring it to a virtual wallet. The last step requires aggregating the data labeled and reporting significant statistics to the system administrator. The payment mechanism is at the core of the labeling system and has a two-way communication relation with the preceding and succeeding steps. That is, a breakdown in communication would result from a failure of the administrator to embed the practical features of the credit function into the body of the labeling system. Moreover, the corresponding function or payment mechanism is crucial for feeding the labeling aggregation and reporting results with the required information.

3.2. Credit Function Payment

We review the original skip-based formula of [10] before describing the modifications proposed to the credit function of the labeling system. This formula is used to define the basic version of our credit function. Table 2 describes the notation used together with the corresponding definitions and mathematical domains.

Table 2. Notations of the payment function.

Symbol	Definition	Domain
N	Total number of questions	$\mathbb{Z}^+$
G	Total number of golden questions	$[0, N] \in \mathbb{Z}^+$
i	Counter of golden questions IDs indexed from 1 through G	$\{1, \dots, G\}$
t	Type of golden items	$\{+, -\}$
s	Possible response of users to golden items	$\{Correct, Incorrect, Skip\}$
$\mu_{\max}$	Maximum credit paid when providing correct answers to all golden questions	$\mathbb{R}^+$
$\mu_{\min}$	Minimum credit paid in terms of users' participation	$[0, \mu_{\max})$
T	Confidence level or shape parameter of the credit function	$(0, 1]$
q_i^s	Response state to the ith golden item	$\begin{cases} 1, \text{ if } ith \text{ golden item is answered via state } s \\ 0, \text{ otherwise} \end{cases}$
λ_i^t	Type of ith golden question	$\begin{cases} 1, \text{ if } ith \text{ golden item is } type\ t \\ 0, \text{ otherwise} \end{cases}$
$Bonus_s^t$	Coefficient assigned to a response from S given to a golden type t question	$\left\{ \begin{array}{l} Bonus^+_{Correct}, Bonus^+_{Skip}, Bonus^+_{Incorrect}, \\ Bonus^-_{Correct}, Bonus^-_{Skip}, Bonus^-_{Incorrect} \end{array} \right\}$
g_s^t	Number of type t golden items assigned state s as answer	$[0, G]$
g_s°	Number of unspecified golden items (according to the original formulation setting) assigned state s as answer	$[0, G]$
$Bonus^\circ_{Correct}$	Coefficient assigned to a correct answer given to an unspecified golden question (according to the original formulation setting)	$\frac{1}{T}$
$Bonus^\circ_{Incorrect}$	Coefficient assigned to an incorrect answer given to an unspecified golden question (according to the original formulation setting)	0
$Bonus^+_{Correct}$	Coefficient assigned to a correct answer given to a positive-type golden question, e.g., a True-sign (✔) assignment to the celebrity image detection	$\frac{1}{T}$
$Bonus^{+,-,\circ}_{Skip}$	Coefficient assigned to skipping any type of golden question	1

Table 2. *Cont.*

Symbol	Definition	Domain
$Bonus^{+}_{Incorrect}$	Coefficient assigned to an incorrect answer given to a positive-type golden question, e.g., a False-sign (×) assignment to the celebrity image detection	$\{0, \{T^x \mid x \in \mathbb{R}^+\}\}$
$Bonus^{-}_{Correct}$	Coefficient assigned to a correct answer given to a negative-type golden question, e.g., a False-sign (×) assignment to the celebrity image detection	$\frac{1}{T}$
$Bonus^{-}_{Incorrect}$	Coefficient assigned to an incorrect answer given to a negative-type golden question, e.g., a True-sign (✔) assignment to the celebrity image detection	$[0, Bonus^{+}_{Incorrect}]$

$$\begin{aligned}&(\mu_{\max}-\mu_{\min})T^{G}\prod_{i=1}^{G}\left(\sum_{s} q_i^s Bonus_s^{\circ}\right)+\mu_{\min}\\ &=(\mu_{\max}-\mu_{\min})T^{G}\left\{\left(Bonus^{\circ}_{Correct}\right)^{g^{\circ}_{Correct}}\left(Bonus^{\circ}_{Incorrect}\right)^{g^{\circ}_{Incorrect}}\left(Bonus^{\circ}_{Skip}\right)^{g^{\circ}_{Skip}}\right\}+\mu_{\min}\end{aligned} \tag{1}$$

The original formulation of the payment function obeys Relation (1), in which the values of the incorrect, skip, and correct coefficients are equal to 0, 1, and $\frac{1}{T}$, respectively. Note that the type of golden questions is omitted from the computation, which results in only three coefficients: $Bonus^{\circ}_{Correct}$, $Bonus^{\circ}_{Incorrect}$, and $Bonus^{\circ}_{Skip}$. An underlying assumption is $\sum_{s} q_i^s = 1, \forall i$, which imposes a single label per question. The extensive form of the formula is completed through the $g^{\circ}_{Correct}$, $g^{\circ}_{Incorrect}$, and g°_{Skip} exponents included within the multiplication of the corresponding coefficients.

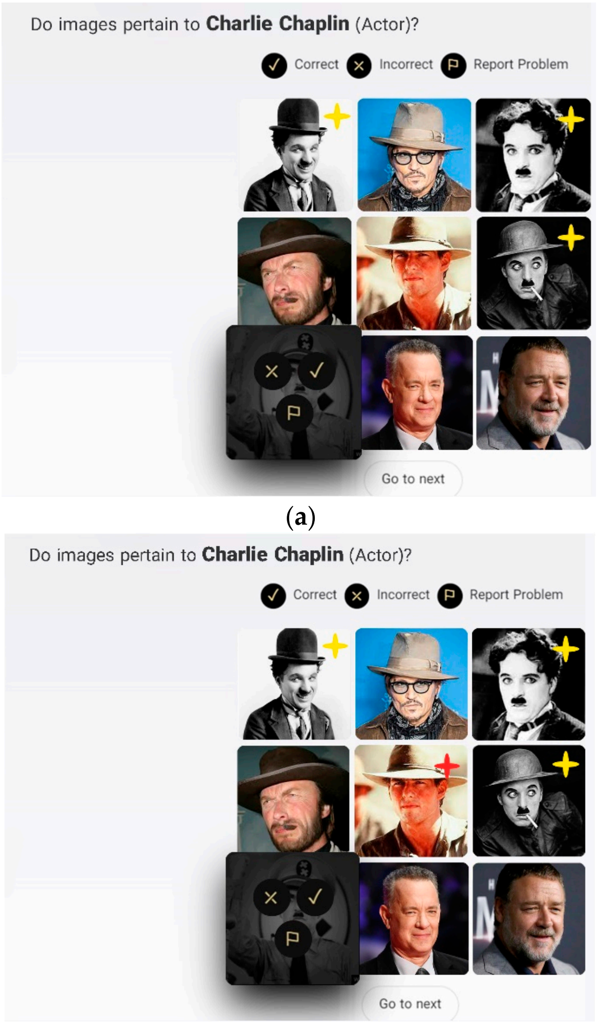
The reasoning behind the choice of $\frac{1}{T}$ is to give the maximum credit to an individual assigning correct labels to all the golden questions. If a contributor responds to all the golden questions correctly, the formula becomes $(\mu_{\max}-\mu_{\min})T^{G}\left(\frac{1}{T}\right)^{G}+\mu_{\min}$, which yields $\mu_{\max}$. Finally, if the respondent skips a choice, a coefficient of 1 is entered into the formula, preserving the credit value intact.

3.2.1. Solution to RQ1 (Preliminary)

If a contributor assigns the wrong label to a single golden question, a zero value is introduced in the formula, resulting in a credit equivalent to $\mu_{\min}$. This is, indeed, the most controversial feature of the original formula, leading to C1. From the perspective of a contributor, it would be discouraging to contribute to a labeling system that returns the minimum credit due to a single error. We provide a milder condition where the default value of $Bonus_{False}$ is substituted by an inverse function of $Bonus_{True}$, i.e., T^x, where x is derived using sensitivity analysis. This modification resolves *RQ1*, though the actual solution will be completed by taking the descriptions of Section 4 into account.

3.2.2. Solution to RQ2

A subsequent problem is associated with the way golden questions are assigned to the dataset items. For instance, consider Figure 2, which illustrates a typical labeling case. There are nine pictures per sheet and users must determine whether they belong to the celebrity named (✔) or not (×). The Report (⚐) and Skip (Go to next) options are also available alternatives. Assume that the administrator of the labeling system sets a specific number of photos as golden items to evaluate the accuracy of the user. In this figure, the star signs attached to some images are symbols of golden items known to the administrator but invisible to ordinary users. Focus now on *C2*. Is it important to define a particular type of golden item? What is the practical consequence of setting all golden items using either True- or False-type answers?



Do images pertain to **Charlie Chaplin** (Actor)?

(**a**)

Do images pertain to **Charlie Chaplin** (Actor)?

(**b**)

Figure 2. True (✔)/False (×)-type labeling questions. (**a**) Considering positive golden items. (**b**) Considering both positive and negative golden items.

At first glance, one may define only True-type golden items since the consistency of a photo with the given celebrity's name is more important than detecting inconsistent items. In other words, consider the question "Are the images related to Charlie Chaplin (Actor)?". Assigning True-type labels to the photos that belong to this actor is more important than attributing False-type labels to the ones which are not Chaplin. In fact, when an image is not related to the celebrity addressed, it may belong to any other individual whose identity is irrelevant to evaluate the machine learning practice. By resorting to such reasoning, the administrator would probably consider True-type golden items validating the actions of spammers. That is, a spammer could simply submit a True sign (✔) for all the questions without even considering their content. Since the golden items correspond to True-type responses and wrong answers to the rest of the images do not affect the payment, a spammer would collect the whole credit while labelling the non-golden items wrongly. To cope with the fraudulent actions of spammers, one straightforward solution involves incorporating golden items with False-type answers. To clearly illustrate the idea, golden items with True- and False-type answers are denoted as positive and negative ones, respectively.

$$
\begin{aligned}
&(\mu_{\max}-\mu_{\min})T^{G}\prod_{i=1}^{G}\left(\sum_{t}\sum_{s}\lambda_i^t q_i^s Bonus_s^t\right)+\mu_{\min}\\
&=(\mu_{\max}-\mu_{\min})T^{G}\left\{\begin{array}{l}\left(Bonus^{+}_{Correct}\right)^{g^{+}_{Correct}}\left(Bonus^{+}_{Incorrect}\right)^{g^{+}_{Incorrect}}\left(Bonus^{+}_{Skip}\right)^{g^{+}_{Skip}}\\ \left(Bonus^{-}_{Correct}\right)^{g^{-}_{Correct}}\left(Bonus^{-}_{Incorrect}\right)^{g^{-}_{Incorrect}}\left(Bonus^{-}_{Skip}\right)^{g^{-}_{Skip}}\end{array}\right\}+\mu_{\min}
\end{aligned}
\tag{2}
$$

In Figure 2a, three golden images (whose pictures include yellow stars) require submitting a True sign. In Figure 2b, a negative-type golden image is also defined, namely, the second image in the second row displaying a red star. In this case, the correct answer implies choosing the False (×) sign. Thus, the whole score cannot be obtained by selecting only True icons. We transform the original formulation into Relation (2) by distinguishing the type of golden items and introducing the type of indicator parameter (i.e., λ_i^t). This expression doubles the number of coefficients relative to the original function, distinguishing between the bonuses assigned to the positive and negative golden items. We allow for different values of the coefficients assigned to the negative golden items. For instance, the administrator could increase the penalty rate of users who submit incorrect answers to negative items and allow it to approach zero, imposing a severe loss on the credit function.

3.2.3. Solution to RQ3 (Preliminary)

Addressing the third concern *C3* involves tuning the shape parameter (T), particularly when facing large-scale datasets. Table 3 displays the trends displayed by the credit values of small- and large-scale instances under different configurations. The data are sorted in terms of correctly responding to a single, half, and the whole number of golden items, absent of any incorrect label ($g^{+,-}_{Incorrect} = 0$). Consider the example of [10], who set $T = 0.5$, for a total of 10 questions ($N = 10$) where 30% are golden ($Q = 3$). Assume now that the maximum and minimum credits equal 80 ($\mu_{\max} = 80$) and 0 ($\mu_{\min} = 0$) units, respectively. Within this framework, the credit obtained by users would equal 20, 40, and 80 after answering 1, 2, and 3 golden questions correctly, respectively. Next, consider a larger dataset with 1000 questions. Applying the same proportions of the previous example, the parameters would be given by $Q = 300$, $\mu_{\max} = 8000$, and $\mu_{\min} = 0$. However, selecting a proportional shape parameter would violate the maximum threshold value of 1.

Table 3. Credit values of small- and large-scale instances.

Parameter Setting	**$Q = 3$, $\mu_{\max} = 80$, $Q = 300$, $\mu_{\max} = 8000$**									
	$T = 0.5$		**$T = 0.5$**		**$T = 0.95$**		**$T = 0.98$**		**$T = 0.985$**	
	$g^{+,-}_{Correct}$	**Value**	$g^{+,-}_{Correct}$	**Value**	$g^{+,-}_{Correct}$	**Value**	$g^{+,-}_{Correct}$	**Value**	$g^{+,-}_{Correct}$	**Value**
$g^{+,-}_{Incorrect} = 0$	**1**	20	**1**	7.85×10^{-87}	**1**	0.002	**1**	19.04	**1**	87.20
	2	40	**150**	5.61×10^{-42}	**150**	3.64	**150**	386.37	**150**	828.94
	3	80	**300**	8000	**300**	8000	**300**	8000	**300**	8000

By keeping the shape parameter value unchanged (e.g., $T = 0.5$), the income obtained by users when correctly submitting 1 and 150 golden items is negligible. There is also a substantial income jump when answering the last golden question correctly. The resulting credit distribution would lessen the reliability of the system from the perspective of users. Most individuals would dismiss the slow credit increments obtained despite responding correctly to a considerable number of questions. An increase in the shape parameter from 0.5 to 0.95 would still deliver credit values that remain insufficient to stimulate users. The output of the credit function becomes more meaningful as T rises to 0.98. After this value, even marginal decimal increments (e.g., 0.985) would significantly increase the outputs. The shape parameter of the formula requires tuning via sensitivity analysis to yield a workable and encouraging system.

However, as the size of the dataset and the number of golden items increases, sensitivity analysis is insufficient to manage the payments in a real labeling system. Let $N = 1000$, $Q = 300$, $\mu_{\max} = 8000$, and $T = 0.985$. Consider the case with 1000 potential contributors. Assume now that users label a single question and leave the system after collecting the credit. How much will the system pay to this type of user? Any of these users facing a golden item by chance will receive 87.2 units when answering correctly. The existence of 300 golden items implies that the total maximum payment equals 26,160, which exceeds the budget assigned to the labeling practice of such a dataset (e.g., $\mu_{\max} + \mu_{\min} = 8000$).

This drawback follows from the fact that the formula does not distribute the total budget linearly among the golden items answered correctly or account for the number of contributors. In addition, the formula does not incorporate a mechanism to avoid violating the budget constraint when multiple individuals complete the labeling of all dataset items. It works well when users label the complete set of questions individually, but this requirement is not feasible with large-scale datasets. The output of the credit function would become inconsistent with the pre-defined budget if the contributors left the system with incomplete labels (i.e., whenever the items labeled are less than N). One solution is to prevent users from receiving any credit until completing all the questions. However, this may be considered an unfair obligation since users may not have enough time to wrap up labeling for large values of N.

Administrators often annotate each item of the dataset with more than two users and aggregate the result to infer the correct labels. At the same time, finding users willing to label the items of a large-scale dataset is far from easy. The proposed simple yet workable solution is to break down the dataset into smaller groups (called targets) and prevent users from collecting their credit prior to completing a target. For instance, assume that administrators require three labels per item. In the example with $N = 1000$, a target of size 100 can be defined, requiring 30 users to label a selected target instead of asking three users to annotate 1000 labels each.

We conclude by noting that the solution to RQ3 will be completed through the analyses performed in Section 4.

3.2.4. Solution to RQ4

The distribution of golden items among ordinary data impacts the output of the payment function and the efficacy of the proposed labeling system. Such distribution should guarantee the inclusion of a pre-defined number of golden items within a target since any additional ones would require overspending the budget of the system. In addition, the order in which golden items are displayed should not be predictable to counteract the actions of potential scammers. The design of an applied distribution mechanism requires choosing between a fixed count of questions for each target and a variable one. In the fixed strategy, the number of images displayed to a user equals the target size. In the variable setting, this number is unknown and images are displayed as long as the user submits True/False answers until the target size is reached. If the administrator follows the fixed strategy, a target size of 100 defines the number of questions shown to the user. The target is terminated whenever he/she submits a True/False/Skip answer to the 100th question. The same target in the variable setting will not be terminated until the user hits 100th True/False answers. In this case, Skip answers are not counted, and questions are displayed until the sum of True and/or False annotations reaches the pre-defined size of the target. To cope with both strategies, straightforward but effective approaches are presented.

Consider the fixed strategy. A division rate for decoupling the questions of each target can be defined as follows: a target of size 40 (including 30 ordinary and 10 golden items) with a division rate of 0.5 is analogous to decomposing the questions into two groups with 20 items each. The contribution of the golden items to each group can be determined through a Bernoulli trial, such as, for instance, mapping 0.2 and 0.8 onto the first and second groups, respectively. The resultant distribution assigns 2 golden items to the first group and 8 golden questions out of the second 20 ones. A final stage of this strategy regards the specification of negative golden questions per target. A percentage of negative golden items can be assigned using a similar intuition as in the general case, e.g., a percentage of 0.2 out of the total would deliver two negative golden items and eight positive ones.

The variable strategy does not limit the number of items shown prior to satisfying the pre-defined target size through True/False labels. This strategy can be implemented by considering a constant number of golden items per sheet (e.g., two golden ones out of the nine images available per webpage). However, such an approach would be vulnerable to budget deficiency since it does not restrict the number of golden items and the credit of a user may go beyond $\mu_{\max}$. If the maximum achievable income is exceeded, the formula can be updated using the pseudo-code described in Relation (3). This equation halts the increasing trend of the function whenever the credit surpasses $\mu_{\max}$. It does so by omitting the coefficient of correct answers. Instead, the current credit (equal to $\mu_{\max}$) of a contributor would be reduced by taking any potential incorrect answers to the golden question into account.

$$\begin{aligned}
&Calculate\ Credit\ as\ usual;\\
&If\ Credit > \mu_{\max}\\
&Credit = \left[\mu_{\max}\left(Bonus^{+}_{Incorrect}\right)^{g^{+}_{Incorrect}}\left(Bonus^{-}_{Incorrect}\right)^{g^{-}_{Incorrect}}\right] + \mu_{\min}
\end{aligned} \tag{3}$$

We ensure the random distribution of golden items through division rates, Bernoulli trials, and equal chance of display in the subgroups defined when implementing a fixed strategy. In the variable strategy, randomness is guaranteed by considering a constant number of golden items with randomized placement per displayed sheet. These methods prevent predictability and align with budget constraints by updating the credit formula dynamically.

4. Sensitivity Analysis (Complementary Solutions to *RQ1* and *RQ3*)

We complete the answers to *RQ1* and *RQ3* by performing a numerical sensitivity analysis of the adjusted credit function. The case-dependent nature of the proposed function prevents us from prescribing a specific formula to derive the corresponding parameters. Instead, we develop a workable procedure to illustrate how the underlying parameters can be tuned. An example is presented to highlight the case-dependent structure of the problem. Further, a sensitivity analysis is performed to provide insights about how to identify the influence of the shape parameter on the function. Capturing such an influence allows us to discuss in detail the penalty coefficient. In this respect, the selection of the penalty coefficient depends on the decision of the system administrator regarding whether or not to choose a strict policy.

We utilize a dataset consisting of 245,000 images of celebrities whose 25% is golden. To aggregate the results, three labels are considered for each image. Thus, the dataset requires 735,000 labels of contributors to be completed, including 183,750 golden items. The total budget equals 15,000,000 units. Thus, we have N = 735,000, Q= 183,750, $\mu_{\max}$ = 15,000,000, and $\mu_{\min} = 0$. As discussed earlier, the formula works accurately when a single user labels the complete set of N items. Clearly, this requirement is not applicable as a prerequisite to receiving the earnings. We therefore segment the dataset into labeling targets of size $N = 500$. The number of golden items and maximum payment defined within each target equal 125 and 10,204, respectively.

To assess the behavior of the credit function, we consider the following key assumptions. First, incorrect labels on golden items are omitted when deriving the credit distribution function. Users are assumed to either submit the correct answer or skip the question. Second, extreme values of the shape parameter are used to identify variations in the shape of the credit function and determine the value of the corresponding coefficients. Finally, to detect output trends under different penalty rates, all labels are initially assumed to be correct, and then the number of incorrect answers is gradually increased.

Figure 3 illustrates the values taken by the credit function for different levels of the shape parameter as the number of correct answers to the golden items increases. To efficiently assess the behavior of the function, the key assumption is to ignore incorrect labels to the golden items. That is, users either submit the correct answer or skip the questions. For example, a value of 100 in the horizontal axis indicates that out of 125 golden items available, 100 labels have been correctly submitted while the other 25 questions were skipped. This figure intuitively describes the influence of T on the distribution shape of the credit function. An inappropriate tuning of T flattens the increasing trend of the credit function, conditioning the subsequent behavior of users. For instance, $T = 0.5$ and $T = 0.999$ represent extreme cases illustrating this feature. When $T = 0.5$, the outputs of the credit function are approximately zero even if 123 out of 125 golden questions have been correctly answered. Note the considerable jump that occurs when the user correctly labels the last two golden items.

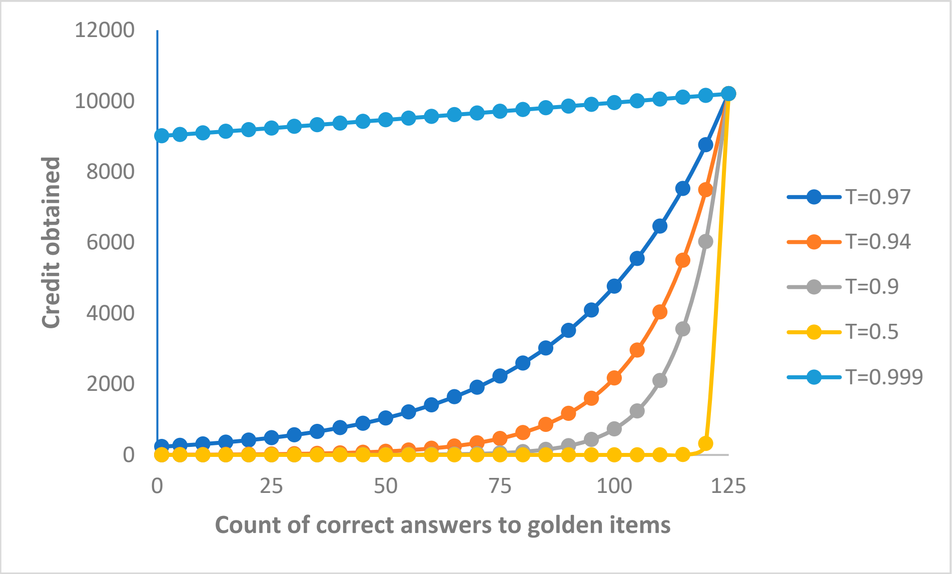


Figure 3. Sensitivity analysis of the shape parameter (T) absent incorrect labeling of golden items.

When $T = 0.999$, the initial output equals 9013, which constitutes a significant percentage of the maximum income, i.e., 10,204. The increment of the function for the remaining items would be considerably slow. Such situations would undermine the trust of users since they could not easily observe their credit progression. Although the final output in all the scenarios proposed equals $\mu_{\max}$, the credit distribution over the count of answers submitted is a crucial incentive for attracting new users. Decision makers should determine the value of T that fits better with their strategy for attracting contributors via trial and error. In this case, $T = 0.97$ provides a reasonable distribution of the credit obtained. Answering correctly to 75 golden items and skipping the remaining 50 would lead to 9706, 2225, 462, 52, and 9.06×10^{-12} units associated with $T = 0.999$, $T = 0.97$, $T = 0.94$, $T = 0.9$, and $T = 0.5$, respectively.

Table 4 describes the distribution of the total budget across $g^{+,-}_{Correct}$ reference points for different values of T. The metric represented defines the ratio of the credit received by a user after submitting a given number of correct answers divided by the total budget available ($\mu_{\max} - \mu_{\min}$). For instance, when $T = 0.94$ and 75 golden items are correctly labeled, we have $\frac{462}{10204}$, leading to the value of 4.53%. The unfair distributions of $T = 0.5$ and $T = 0.999$ are clearly observable in this table, with users being respectively paid nothing (0) and a large percentage of $\mu_{\max}$ (90.48) after submitting 25 correct answers and no incorrect label.

Table 4. Distribution analysis of the credit function with different values of T.

$g^{+,-}_{Correct}$	***T***				
	0.5	**0.9**	**0.94**	**0.97**	**0.999**
	Distribution Rate of $\mu_{\max}$ (%)				
0	0	0	0.04	2.27	88.24
25	0	0	0.21	4.76	90.48
50	0	0.04	0.97	10.18	92.77
75	0	0.52	4.53	21.81	95.12
100	0	7.18	21.29	46.7	97.53
125	100	100	100	100	100

Consider now the incorrect answers to golden items, i.e., $Bonus_{Incorrect}$, described in Figure 4. The horizontal axis displays the count of wrong labels assigned to golden items with the remaining ones assumed to be correct, that is, the number 100 corresponds to the case where 25 and 100 golden items are correctly and incorrectly labeled, respectively. The figure illustrates the relationship between the count of incorrect labels and the credit function when considering a variety of values for $Bonus_{Incorrect}$. This coefficient determines the behavior of the function by replacing the value of zero with T^x. In this example, x is assigned the values 0.2, 0.5, 1, 2, and 3, with T = 0.97. Note that the default setting of the original formulation is given by $Bonus_{Incorrect} = 0$. In this case, the credit function drops to its lowest level right after the user submits a wrong answer to any of the golden items.

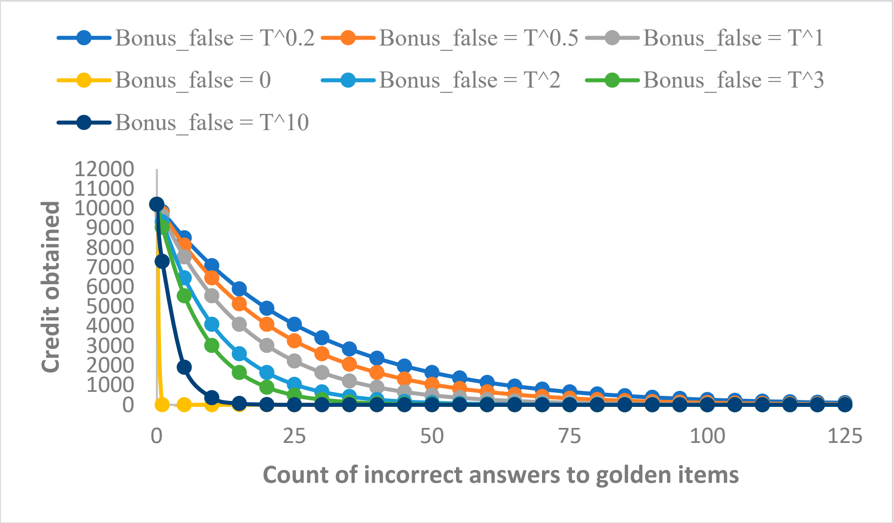


Figure 4. Sensitivity analysis of the shape parameter (T) without skipping any golden item.

Clearly, when all labels are assigned correctly, the credit function hits the ceiling of 10,204. The strict requirements of the original formula can be smoothed by introducing counterpart values through the inverse form of $Bonus_{correct}$, namely, the positive power of T. As the number of incorrect answers approaches 0, the function converges to the default setting while divergencies increase with the number of incorrect answers. Consider the case with 25 incorrect labels and 100 correct ones. When $Bonus_{Incorrect}$ equals 0, 0.994, 0.985, 0.97, 0.941, 0.913, and 0.737, outputs are given by 0, 4091.88, 3256.2, 2225.14, 1039.08, 485.23, and 2.35, respectively. The data labeling system administrator must select the appropriate penalty, ranging from a strict $Bonus_{Incorrect} = T^{10}$, to a more lenient one, $Bonus_{Incorrect} = T^{0.2}$.

We provide additional intuition by defining a metric that measures penalty intensity. Assume that the credit function is independent of the number of wrong answers by adjusting $Bonus_{Incorrect} = 1$. The penalty intensity ratio described in Table 5 is determined by the relative difference between the independent credit and the one received. For instance, consider the case with 25 wrong and 100 right answers. Absent of any penalty, i.e., with $Bonus_{Incorrect} = 1$, the credit function equals 4765.01. When the credit received is based on $Bonus_{Incorrect} = 0.97$, which yields 2225.14, the penalty intensity ratio is given by $\frac{(4765.01-2225.14)}{4765.01}$, that is, 0.533. That is, introducing $Bonus_{Incorrect} = 0.97$ leads to a 53.3% decrease in the credit of users.

Table 5. Relative penalty rate relative to the independent incorrect label case.

$g^{+,-}_{Incorrect}$	$Bonus_{Incorrect}$						
	0	$T^{0.2}$	$T^{0.5}$	T	T^2	T^3	T^{10}
	Relative Penalty Rate (%)						
0	0	0	0	0	0	0	0
25	100	14.13	31.66	53.3	78.19	89.82	99.95
50	100	26.26	53.3	78.19	95.24	98.96	100
75	100	36.67	68.09	89.82	98.96	99.89	100
100	100	45.26	78.19	95.24	99.77	99.99	100
125	100	53.3	85.1	97.78	99.95	100	100

Despite the reliability illustrated through the sensitivity analysis, a simple typo could still lead to setting an inappropriate value for T. To counteract this possibility, we recommend performing a small scale test prior to starting the Go-live labeling of a new dataset. A group of pre-selected reliable users can be enrolled to validate whether the tunned parameters of the credit function, particularly T, work well or not. After the accuracy of the credit function has been validated by the system administrator, the Go-live process of the labeling associated with a new dataset can be initiated.

5. Aggregating Results (Solution to RQ5)

RQ5 involves retrieving the outputs of the labeling system and delivering managerial reports. In this regard, the ultimate goal of the system consists in providing feedback on the outputs obtained from a labeling practice. This process encompasses the following tasks: aggregating the responses received per item, measuring the reliability of answers, assessing the volume of remaining work, and validating the accuracy of golden items as well as the reports describing the financial performance of users.

The aggregation of labeling results will be tackled by separating the analyses related to ordinary and golden items. The answers to ordinary items are unknown to the administrator while golden questions are assigned default labels. Therefore, the analysis of ordinary items should aim at deriving dominant responses under different aggregation rules and reliability metrics. The reason for double checking the responses to golden items is to evaluate the accuracy of their default labels. The answers to golden items are meticulously defined under the supervision of the administrator. Despite this fact, incorrect default labels could still be assigned. Tables 6 and 7 provide a set of metrics for extracting information regarding the status of annotated items within a labeling system.

Table 6. Proposed metrics to evaluate the status of ordinary items.

Status of Ordinary Items	Metric
Complete vs. incomplete items	• Contribution of items with complete labels: Completion $(\%) = \frac{Count\ of\ complete\ ordinary\ items}{Total\ number\ of\ ordinary\ items} \times 100$ • Contribution of items with incomplete labels: Incompletion $(\%) = \frac{Count\ of\ incomplete\ ordinary\ items}{Total\ number\ of\ ordinary\ items} \times 100$

Table 6. *Cont.*

Status of Ordinary Items	Metric
Complete items	• Contribution of complete items with dominant labels: $\text{Domination } (\%) = \frac{Count\ of\ complete\ ordinary\ items\ with\ dominant\ result}{Total\ number\ of\ complete\ ordinary\ items} \times 100$ • Contribution of complete items with semi-dominant labels: $\text{Semi} - \text{domination } (\%) = \frac{Count\ of\ complete\ ordinary\ items\ with\ semi-dominant\ result}{Total\ number\ of\ complete\ ordinary\ items} \times 100$ • Contribution of complete items with non-dominant labels: $\text{Non} - \text{domination } (\%) = \frac{Count\ of\ complete\ ordinary\ items\ with\ non-dominant\ result}{Total\ number\ of\ complete\ ordinary\ items} \times 100$
Complete items with dominant result	• Contribution of complete dominant items with True response: $\text{True} - \text{based domination } (\%) = \frac{Count\ of\ complete\ dominant\ ordinary\ items\ with\ True\ response}{Total\ number\ of\ complete\ dominant\ ordinary\ items} \times 100$ • Contribution of complete dominant items with False response: $\text{False} - \text{based domination } (\%) = \frac{Count\ of\ complete\ dominant\ ordinary\ items\ with\ False\ response}{Total\ number\ of\ complete\ dominant\ ordinary\ items} \times 100$ • Contribution of complete dominant items with Report response: $\text{Report} - \text{based domination } (\%) = \frac{Count\ of\ complete\ dominant\ ordinary\ items\ with\ Report\ response}{Total\ number\ of\ complete\ dominant\ ordinary\ items} \times 100$
Incomplete items	• Contribution of incomplete items with no label: $\text{No label incompletion } (\%) = \frac{Count\ of\ incomplete\ ordinary\ items\ with\ no\ label}{Total\ number\ of\ incomplete\ ordinary\ items} \times 100$ • Contribution of incomplete items with at least one single label: $\text{Min label incompletion } (\%) = \frac{Count\ of\ incomplete\ ordinary\ items\ with\ at\ least\ one\ single\ label}{Total\ number\ of\ incomplete\ ordinary\ items} \times 100$
Accuracy of user responses	• Contribution of users with high accuracy rate in labeling complete dominant items: $\text{High} - \text{level users contribution } (\%) = \frac{Count\ of\ completed\ dominant\ ordinary\ items\ by\ highly\ accurate\ users}{Total\ number\ of\ complete\ dominant\ ordinary\ items} \times 100$ • Contribution of users with medium accuracy rate in labeling complete dominant items: $\text{Medium} - \text{level users contribution } (\%) =$ $\frac{Count\ of\ completed\ dominant\ ordinary\ items\ by\ medium\ accurate\ users}{Total\ number\ of\ complete\ dominant\ ordinary\ items} \times 100$ • Contribution of users with low accuracy rate in labeling complete dominant items: $\text{Low} - \text{level users contribution } (\%) = \frac{Count\ of\ completed\ dominant\ ordinary\ items\ by\ lowly\ accurate\ users}{Total\ number\ of\ complete\ dominant\ ordinary\ items} \times 100$ *The accuracy rate of each user is computed as follows:* $\text{Accuray rate of User } (\%) = \frac{\text{Count of golden items correctly labelled}}{\text{Total number of golden items labelled}} \times 100$ $\text{Average accuray rate of Users that labeled a dominant item } (\%) =$ $\frac{Sum\ of\ accuracy\ rates\ of\ Users\ that\ labeled\ a\ dominant\ item}{Total\ number\ of\ Users\ labelling\ a\ dominant\ item} \times 100$

The status of ordinary items can be analyzed using different metrics. It is crucial to assess the percentage of ordinary data that have received the required number of labels, defined as complete items. This implies differentiating between the contribution of items with dominant, semi-dominant, and non-dominated results. If there is a threshold of three labels per item, three synchronous answers (e.g., True or False) constitute a dominant result, showing a strong consensus of users over the corresponding item. In the case of two similar labels (e.g., True) and one opposite label (e.g., False), the system assigns a semi-dominant state to the item. The non-dominant status occurs when the three users report the item or provide distinct responses (i.e., True, False, and Report). The complete items with a dominant result will display True, False, or Report labels. Among the incomplete ones, the rate of unlabeled items, as well as cases with at least a single answer submitted, will convey useful information to the admin.

Table 7. Proposed metrics to evaluate the status of golden items.

Status of Golden Items	Metric
Consistency level of responses	• Contribution of items with a high level of consistency: $\text{High level of consistency } (\%) = \frac{Count\ of\ golden\ items\ labeled\ with\ a\ high\ level\ of\ consistency}{Total\ number\ of\ golden\ items} \times 100$ • Contribution of items with a medium level of consistency: $\text{Medium level of consistency } (\%) = \frac{Count\ of\ golden\ items\ labeled\ with\ a\ medium\ level\ of\ consistency}{Total\ number\ of\ golden\ items} \times 100$ • Contribution of items with a low level of consistency: $\text{Low level of consistency } (\%) = \frac{Count\ of\ golden\ items\ labeled\ with\ a\ low\ level\ of\ consistency}{Total\ number\ of\ golden\ items} \times 100$ *The consistency level of each golden item is computed as follows:* $\text{Consistency level of golden item } (\%) = \frac{Count\ of\ answers\ consistent\ with\ the\ predefined\ label\ of\ the\ golden\ item}{Total\ number\ of\ answers\ submitted\ to\ the\ golden\ item}$
Percentage of responses' type	• Contribution of golden responses among all labeled items: $\text{Golden labels } (\%) = \frac{Count\ of\ golden\ items\ labeled}{Total\ number\ of\ items\ labeled} \times 100$ • Contribution of non-golden (ordinary) responses among all labeled items: $\text{Non} - \text{golden labels } (\%) = \frac{Count\ of\ non{-}golden\ items\ labeled}{Total\ number\ of\ items\ labeled} \times 100$

We must also define a reliability criterion for the aggregated data. The criterion proposed maps the accuracy of the responses of contributors to golden data into the complete dominant items. In particular, it specifies the contribution percentage of highly, moderately, and lowly accurate users in the answers submitted to dominant items. A highly accurate quality is attributed to a user whose number of correct responses associated with the golden items divided by the total number of golden items labeled is greater than or equal to 80%. When this amount is between 50% and 80%, the corresponding user is assigned a medium level of accuracy while a percentage below 50% implies that the system faces a low accurate user. Note that the threshold values may vary from case to case and should be set based on the system administrator criteria and preferences. The system must therefore allow for the administrator to contable the threshold values accordingly.

We must note that although our approach is primarily designed to dissuade spammers from participating, a specific strategy is implemented based on our reliability criterion. In particular, the system administrator should set a threshold for the minimum acceptable amount of the average accuracy rate of users who labeled a dominant item. A low value of this criterion is, indeed, the prospective outcome of spammer (or, equivalently, careless user) activity. As the value of the criterion falls below the threshold, the system changes the status of the completed item to incomplete by omitting the labels of the corresponding spammers. The item can be further completed through the contribution of other users

The participation rate of each user type per complete dominant item constitutes a reliability indicator for the system administrator. As the portion of dominant items with highly accurate users increases, the overall performance of the labeling practice becomes more promising. Furthermore, the consistency level of the golden items demonstrates whether they have been labeled correctly. If more than 80% of the answers submitted to a golden item correspond with its pre-determined label, the consistency level is regarded as high. In the same vein, if this number is between 50% and 80%, or lower than 50%, the consistency levels are defined as medium or low, respectively. The distribution of the items answered between golden and non-golden is yet another metric that can be used to analyze the output of the system.

Given the key contribution of output objects to machine learning practice, assessing the overall accuracy of labels across the entire dataset may not provide a meaningful measure of performance. In our framework of analysis, the influence of each labeled item may vary significantly. Mislabeled items can have negative ramifications and lead to excessive processing costs depending on the context and specific machine learning application. We have therefore proposed a reliability criterion to analyze the quality of labels on an item-by-item basis. This approach ensures that the output of each completed item can be identified and categorized as low, medium, or high quality. If the administrator is not satisfied with the quality level of certain items, the system allows for the corresponding items to receive further labels until the desired quality is reached. In this way, overall improvements are guaranteed for the labeled dataset by focusing on the individual measurement of the accuracy of the items' labels.

Table 8 describes the labeling result of a hypothetical dataset including 10 images. The images are listed i_1 through i_{10} where the second (i_2^{*+}), fifth (i_5^{*+}), and eighth (i_8^{*-}) ones are positive, positive, and negative golden types, respectively. In this case, the correct label of the positive (negative) golden items is True (False). A, B, and C are three users contributing to the labeling practice. The possible responses are True (T), False (F), and Report (R) with three labels required per item. If doubtful about the correct label, users can skip (S) the images. For instance, user A has skipped the image while both B and C have assigned it a True label.

Table 8. Hypothetical labeling result with 10 items and the contribution of 3 users.

User	Item									
	i_1	i_2^{*+}	i_3	i_4	i_5^{*+}	i_6	i_7	i_8^{*-}	i_9	i_{10}
A	T	T	T	F	T	F	R	F	S	S
B	T	T	F	F	T	F	T	T	S	T
C	T	S	F	F	F	F	F	F	S	T

Table 9 graphically and numerically represents the metrics related to the ordinary items. For instance, to compute the *No label incompletion metric*, the denominator enumerates the items that do not receive the three pre-defined labels, i.e., two items, including i_9 and i_{10}. Actually, i has no labels assigned since all users have skipped it and i_{10} is incomplete since it has only been labeled by B and C and skipped by A. The numerator is defined by the number of items that do not have a label, i.e., item i_9. Thus, the *No label incompleteness rate* is $\frac{1}{2}$, which means that 50% of the incomplete items have no label.

Table 9. Metrics of the ordinary data associated with the hypothetical study case.

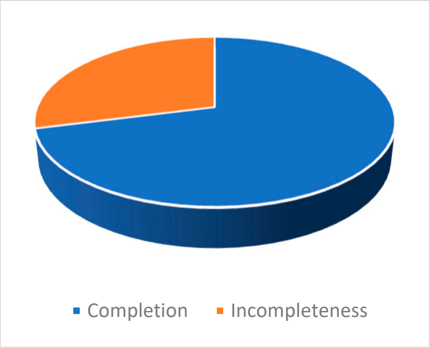
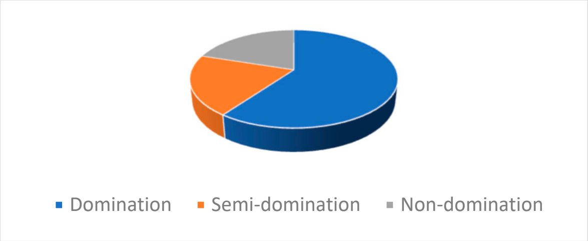
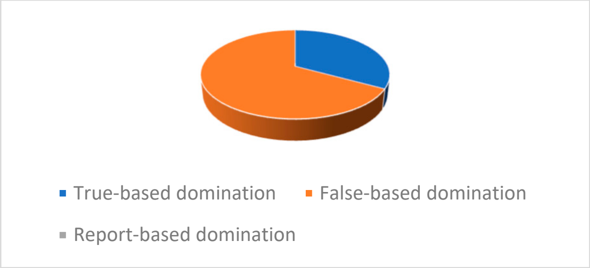
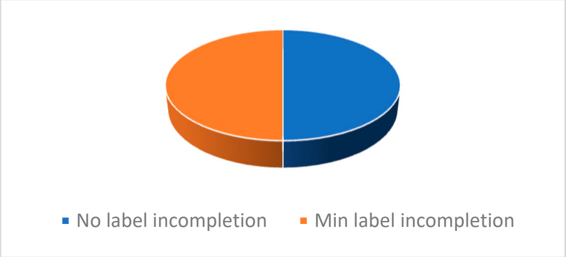
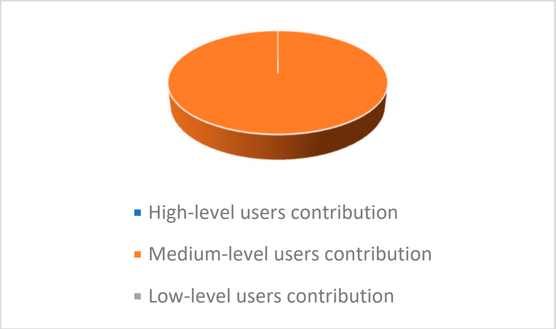
Graphical Representation	Computation
	$\text{Completion }(\%) = \frac{\|i_1,i_3,i_4,i_6,i_7\|}{\|i_1,i_3,i_4,i_6,i_7,i_9,i_{10}\|} = \frac{5}{7} \times 100 = 71.43\%$ $\text{Incompleteness }(\%) = \frac{\|i_9,i_{10}\|}{\|i_1,i_3,i_4,i_6,i_7,i_9,i_{10}\|} = \frac{2}{7} \times 100 = 28.57\%$
	$\text{Domination }(\%) = \frac{\|i_1,i_4,i_6\|}{\|i_1,i_3,i_4,i_6,i_7\|} = \frac{3}{5} \times 100 = 60\%$ $\text{Semi} - \text{domination }(\%) = \frac{\|i_3\|}{\|i_1,i_3,i_4,i_6,i_7\|} = \frac{1}{5} \times 100 = 20\%$ $\text{Non} - \text{domination }(\%) = \frac{\|i_7\|}{\|i_1,i_3,i_4,i_6,i_7\|} = \frac{1}{5} \times 100 = 20\%$

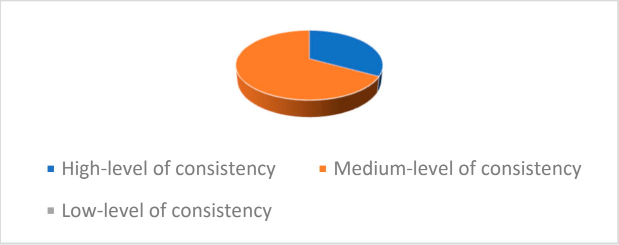
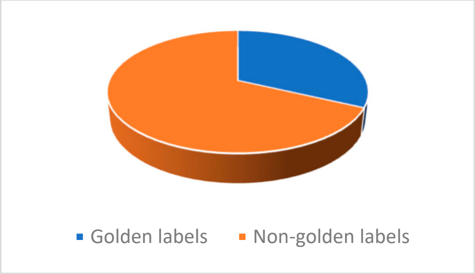
Table 9. *Cont.*

Graphical Representation	Computation
	$\text{True} - \text{based domination } (\%) = \frac{\lvert i_1 \rvert}{\lvert i_1, i_4, i_6 \rvert} = \frac{1}{3} \times 100 = 33.33\%$ $\text{False} - \text{based domination } (\%) = \frac{\lvert i_4, i_6 \rvert}{\lvert i_1, i_4, i_6 \rvert} = \frac{2}{3} \times 100 = 66.67\%$ $\text{Report} - \text{based domination } (\%) = \frac{\lvert \varnothing \rvert}{\lvert i_1, i_4, i_6 \rvert} = 0 \times 100 = 0\%$
	$\text{No label incompletion } (\%) = \frac{\lvert i_9 \rvert}{\lvert i_9, i_{10} \rvert} = \frac{1}{2} \times 100 = 50\%$ $\text{Min label incompletion } (\%) = \frac{\lvert i_{10} \rvert}{\lvert i_9, i_{10} \rvert} = \frac{1}{2} \times 100 = 50\% \times 100$
	$\text{Accuray rate of A } (\%) = \frac{\lvert i_2^{*+}, i_5^{*+}, i_8^{*-} \rvert}{\lvert i_2^{*+}, i_5^{*+}, i_8^{*-} \rvert} = \frac{3}{3} = 100\%$ $\text{B } (\%) = \frac{\lvert i_2^{*+}, i_5^{*+} \rvert}{\lvert i_2^{*+}, i_5^{*+}, i_8^{*-} \rvert} = \frac{2}{3} = 66.67\%;\ \text{C } (\%) = \frac{\lvert i_8^{*-} \rvert}{\lvert i_5^{*+}, i_8^{*-} \rvert} = \frac{1}{2} = 50\%$ $\text{Average accuray rate of users who labelled } i_1, i_4, i_6 \ (\%) = \frac{100+66.67+50}{3} = 72.22\%$ $\text{High} - \text{level users contribution } (\%) = \frac{\lvert \varnothing \rvert}{\lvert i_1, i_4, i_6 \rvert} = \frac{0}{3} \times 100 = 0\%$ $\text{Medium} - \text{level users contribution } (\%) = \frac{\lvert i_1, i_4, i_6 \rvert}{\lvert i_1, i_4, i_6 \rvert} = \frac{3}{3} \times 100 = 100\%$ $\text{Low} - \text{level users contribution } (\%) = \frac{\lvert \varnothing \rvert}{\lvert i_1, i_4, i_6 \rvert} = \frac{0}{3} \times 100 = 0\%$

This result also implies that dominant items have been completed based on the accuracy of the responses of medium-level users to the golden data. That is, dominant complete items i_1, i_4, and i_4 are labeled by a group of users whose average responses to the golden data range between 50% and 80%. User A has responded correctly to all golden data, obtaining the whole accuracy score. User B has provided a wrong answer to i_8^{*-}, leading to two right responses out of the three golden ones and an accuracy rate of 66.67%. User C has only labeled two golden items and the response to i_5^{*+} is incorrect, resulting in an accuracy rate of 50%. The average rate of users who contribute to labeling the complete dominant items equals 72.22%. This value implies that 100% of the dominant complete items have been labeled by users with a medium accuracy level.

Table 10 describes the results derived relative to the golden items. To determine the consistency level of the golden data, we must first calculate the fraction of users providing consistent responses to the default positive and negative labels. For instance, two users have labeled i_2^{*+} via correct True responses. Hence, the responses of users are fully consistent with the default positive label of i_2^{*+}, resulting in a 100% consistency level. Similarly, the corresponding level for both i_5^{*+} and i_8^{*+} is 66.67%, consistent with their default labels. The report delivered to the system administrator should state that 33.33% and 66.67% of the golden items involve a high and medium level of consistency, respectively. Note that there is no golden item with a consistency level below 50%. Moreover, out of 25 labels registered, 17 of them correspond to ordinary items and the 8 remaining answers are allocated to the golden items. In other words, 32% of the labeling is devoted to golden items whereas ordinary data account for 68%.

Table 10. Metrics of the golden data associated with the hypothetical study case.

Graphical Representation	Computation
	$\text{Consistency level of } i_2^* \ (\%) = \frac{\lvert A,B \rvert}{\lvert A,B \rvert} = \frac{2}{2} = 100\%$ $i_5^*(\%) = \frac{\lvert A,B \rvert}{\lvert A,B,C \rvert} = \frac{2}{3} = 66.67\%;\ i_8^*(\%) = \frac{\lvert A,C \rvert}{\lvert A,B,C \rvert} = \frac{2}{3} = 66.67\%$ $\text{High level of consistency } (\%) = \frac{\lvert i_2^{*+} \rvert}{\lvert i_2^{*+}, i_5^{*+}, i_8^{*-} \rvert} = 33.33\%$ $\text{Medium level of consistency } (\%) = \frac{\lvert i_5^{*+}, i_8^{*-} \rvert}{\lvert i_2^{*+}, i_5^{*+}, i_8^{*-} \rvert} = \frac{2}{3} = 66.67\%$ $\text{Low level of consistency } (\%) = \frac{\lvert \varnothing \rvert}{\lvert i_2^{*+}, i_5^{*+}, i_8^{*-} \rvert} = \frac{0}{3} = 0\%$
	$\text{Golden labels } (\%) = \frac{8}{25} = 32\%$ $\text{Non} - \text{golden labels } (\%) = \frac{17}{25} = 68\%$

The functionality of the metrics proposed in our system can be compared in detail with those of recent data labeling systems and algorithms. The labeling system designed by [31] introduces a reliability metric associated with the performance of users—while ours provides a bridge between the latter and the quality of the data labeled. These authors also set a minimum reliability threshold below which the responses of the corresponding users are omitted from the final aggregated results. Ref. [32] designed an accuracy metric to evaluate the output results obtained. Their metric was defined in terms of the number of tasks (items) whose estimated labels (aggregated result) were consistent with their True labels (golden ones) divided by the total number of tasks. This definition implies that the accuracy metric in their study does not lead to an entity-based output. The corresponding accuracy metric is, indeed, applied to justify the entire dataset. Conversely, as discussed above, our paper proposes an item-by-item evaluation of accuracy, highlighting its entity-based nature. The data labeling system presented by [16] introduces a metric called precision that specifically measures the accuracy of individual annotations against a gold standard. They also define an F1 score, which is a summary statistic that provides an overall measure of performance but does not offer detailed insights into specific aspects of the labeling process. In contrast, our metrics provide a more granular analysis of performance in light of the labeling process.

The metrics proposed provide a comprehensive overview of the labeling process, ensuring informed decision making. By aggregating the responses received per item, the administrator can evaluate the consensus existing among users and identify items needing further attention. Being able to measure the reliability of answers allows the administrator to evaluate the accuracy and consistency of user contributions, identify high-performing users and address potential problems with less reliable contributors. Assessing the volume of remaining work enables efficient resource allocation and prioritization of tasks. Additionally, validating the accuracy of golden items and analyzing reports on financial performance ensures that the labeling system maintains high standards and aligns with organizational goals. This comprehensive feedback loop helps administrators optimize the labeling process, enhance data quality, and ultimately improve overall system performance.

While general accuracy metrics focus solely on the correctness of labels, the proposed metrics also consider the completeness and dominance of responses, the reliability of contributors, and the consistency of golden items. This holistic approach ensures that not only is the accuracy of individual labels assessed, but also the overall reliability and robustness of the labeling process. Finally, aggregating responses per item and distinguishing between ordinary and golden items allow for the identification of patterns and discrepancies that simple accuracy measures might miss.

6. Conclusions

This paper has elaborated on the payment mechanism and reporting framework of data labeling systems. To adopt a workable payment mechanism, we focused on customizing one of the simplest yet most reliable methods in the literature, namely, the skipped-based golden-oriented function. We showed how its rigorous penalty scheme could be moderated by substituting the coefficient of zero with a power function. The behavior of the function was studied numerically, and a sensitivity analysis performed to tune its parameters. The value of the shape parameter was selected through two metrics defined to account for the allocation of credit and intensity of penalties. Negative golden items were introduced to hedge against the credit increase in spammers and careless users.

The distribution of golden data was used to illustrate how the enumeration of Skip labels could negatively influence the interaction of users with the system. The aggregation of results was addressed by configuring a reporting framework using multiple metrics. These metrics were proposed to signal the completion, domination, and consistency status of golden and ordinary items as well as the accuracy of the labels submitted. The quality of the labels was assessed by ranking the performance of users and calculating their contribution to completely labeling the items. Finally, the default values of golden data were double checked for consistency and the proportion of labeled golden versus ordinary items was also analyzed.

Among the potential extensions of this study, software engineering-oriented practices could be defined to develop the labeling system through data models, pseudo-code, and the relationships arising across the tables of the database. As discussed throughout the paper, a cornerstone of our study focused on enhancing the incentives of users to trust the newly launched data labeling system. In this regard, surveys can be carried out to assess the importance that the satisfaction of users has for boosting the corresponding system as well as providing high-quality labels.

Author Contributions: Conceptualization, V.H. and S.J.; methodology, V.H., S.J., F.J.S.-A., S.V.N. and D.D.C.; software, V.H. and S.J.; validation, F.J.S.-A. and S.V.N.; formal analysis, V.H. and S.J.; investigation, F.J.S.-A., S.V.N. and D.D.C.; resources, V.H. and S.J.; data curation, V.H. and S.J.; writing—original draft preparation, V.H., S.J., F.J.S.-A., S.V.N. and D.D.C.; writing—review and editing, V.H., S.J., F.J.S.-A., S.V.N. and D.D.C.; visualization, F.J.S.-A., S.V.N. and D.D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Vahid Hajipour and Sajjad Jalali were both employed by FANAP Co. located in Tehran, Iran. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Thuan, N.H.; Antunes, P.; Johnstone, D. Factors influencing the decision to crowdsource: A systematic literature review. *Inf. Syst. Front.* **2016**, *18*, 47–68. [CrossRef]
2. Bhatti, S.S.; Gao, X.; Chen, G. General framework, opportunities and challenges for crowdsourcing techniques: A Comprehensive survey. *J. Syst. Softw.* **2020**, *167*, 110611. [CrossRef]
3. Yaiprasert, C.; Hidayanto, A.N. AI-driven ensemble three machine learning to enhance digital marketing strategies in the food delivery business. *Intell. Syst. Appl.* **2023**, *18*, 200235. [CrossRef]
4. Lyu, M.; Li, X.; Chen, C.H. Achieving Knowledge-as-a-Service in IIoT-driven smart manufacturing: A crowdsourcing-based continuous enrichment method for Industrial Knowledge Graph. *Adv. Eng. Inform.* **2022**, *51*, 101494. [CrossRef]
5. Majava, J.; Hyvärinen, K. Crowdsourcing-based business model for online customer service: A case study. *Int. J. Value Chain Manag.* **2022**, *13*, 33–46. [CrossRef]

6. Ye, C.; Wang, H.; Lu, W.; Li, J. Effective Bayesian-network-based missing value imputation enhanced by crowdsourcing. *Knowl. -Based Syst.* **2020**, *190*, 105199. [CrossRef]
7. Paullada, A.; Raji, I.D.; Bender, E.M.; Denton, E.; Hanna, A. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* **2021**, *2*, 100336. [CrossRef]
8. Sarı, A.; Tosun, A.; Alptekin, G.I. A systematic literature review on crowdsourcing in software engineering. *J. Syst. Softw.* **2019**, *153*, 200–219. [CrossRef]
9. Shah, N.B.; Zhou, D. Approval Voting and Incentives in Crowdsourcing. *ACM Trans. Econ. Comput.* **2020**, *8*, 13. [CrossRef]
10. Shah, N.B.; Zhou, D. Double or nothing: Multiplicative incentive mechanisms for Crowdsourcing. *J. Mach. Learn. Res.* **2016**, *17*, 1–52.
11. Ghezzi, A.; Gabelloni, D.; Martini, A.; Natalicchio, A. Crowdsourcing: A Review and Suggestions for Future Research. *Int. J. Manag. Rev.* **2018**, *20*, 343–363. [CrossRef]
12. Wang, Y.; Liao, P.C.; Zhang, C.; Ren, Y.; Sun, X.; Tang, P. Crowdsourced reliable labeling of safety-rule violations on images of complex construction scenes for advanced vision-based workplace safety. *Adv. Eng. Inform.* **2019**, *42*, 101001. [CrossRef]
13. Vryzas, N.; Vrysis, L.; Kotsakis, R.; Dimoulas, C. A web crowdsourcing framework for transfer learning and personalized Speech Emotion Recognition. *Mach. Learn. Appl.* **2021**, *6*, 100132. [CrossRef]
14. Khajwal, A.B.; Noshadravan, A. An uncertainty-aware framework for reliable disaster damage assessment via crowdsourcing. *Int. J. Disaster Risk Reduct.* **2021**, *55*, 102110. [CrossRef]
15. He, X.; Zhang, H.; Yang, X.; Guo, Y.; Bian, J. STAT: A Web-based Semantic Text Annotation Tool to Assist Building Mental Health Knowledge Base. In Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI), Xi'an, China, 10–13 June 2019; pp. 1–4.
16. He, X.; Zhang, H.; Bian, J. User-centered design of a web-based crowdsourcing-integrated semantic text annotation tool for building a mental health knowledge base. *J. Biomed. Inform.* **2020**, *110*, 103571. [CrossRef] [PubMed]
17. Vasconcelos, L.; Zahn, J.; Trevisan, D.; Viterbo, J. Engagement by Design Cards: A tool to involve designers and non-experts in the design of crowdsourcing initiatives. *Int. J. Hum.-Comput. Stud.* **2024**, *182*, 103166. [CrossRef]
18. Morschheuser, B.; Hamari, J.; Koivisto, J.; Maedche, A. Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *Int. J. Hum.-Comput. Stud.* **2017**, *106*, 26–43. [CrossRef]
19. Hirth, M.; Hoßfeld, T.; Tran-Gia, P. Cost-Optimal Validation Mechanisms and Cheat-Detection for Crowdsourcing Platforms. In Proceedings of the 2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Seoul, Republic of Korea, 30 June–2 July 2011; pp. 316–321.
20. Wu, W.; Wang, W.; Li, M.; Wang, J.; Fang, X.; Jiang, Y.; Luo, J. Incentive Mechanism Design to Meet Task Criteria in Crowdsourcing: How to Determine Your Budget. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 502–516. [CrossRef]
21. Khan, A.R.; Garcia-Molina, H. CrowdDQS: Dynamic Question Selection in Crowdsourcing Systems. In Proceedings of the 2017 ACM International Conference on Management of Data, Chicago, IL, USA, 14–19 May 2017; pp. 1447–1462.
22. Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; Allahbakhsh, M. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.* **2018**, *51*, 7. [CrossRef]
23. Gadiraju, U.; Kawase, R.; Dietze, S.; Demartini, G. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea, 18–23 April 2015; pp. 1631–1640.
24. Zhang, J.; Wu, X.; Sheng, V.S. Learning from crowdsourced labeled data: A survey. *Artif. Intell. Rev.* **2016**, *46*, 543–576. [CrossRef]
25. Dow, S.; Kulkarni, A.; Klemmer, S.; Hartmann, B. Shepherding the crowd yields better work. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, Seattle, DC, USA, 11–15 February 2012; pp. 1013–1022.
26. Zhang, J.; Sheng, V.S.; Wu, J. Crowdsourced Label Aggregation Using Bilayer Collaborative Clustering. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3172–3185. [CrossRef] [PubMed]
27. Baslyman, M.; Amyot, D.; Mylopoulos, J. Reasoning about Confidence in Goal Satisfaction. *Algorithms* **2022**, *15*, 343. [CrossRef]
28. Feng, Z. IoT data sharing technology based on blockchain and federated learning algorithms. *Intell. Syst. Appl.* **2024**, *22*, 200359. [CrossRef]
29. Tavana, M.; Hajipour, V.; Oveisi, S. IoT-based enterprise resource planning: Challenges, open issues, applications, architecture, and future research directions. *Internet Things* **2020**, *11*, 100262. [CrossRef]
30. Stoykova, S.; Shakev, N. Artificial Intelligence for Management Information Systems: Opportunities, Challenges, and Future Directions. *Algorithms* **2023**, *16*, 357. [CrossRef]
31. Bastanfard, A.; Shahabipour, M.; Amirkhani, D. Crowdsourcing of labeling image objects: An online gamification application for data collection. *Multimed. Tools Appl.* **2024**, *83*, 20827–20860. [CrossRef]
32. Wu, G.; Zhou, L.; Xia, J.; Li, L.; Bao, X.; Wu, X. Crowdsourcing truth inference based on label confidence clustering. *ACM Trans. Knowl. Discov. Data* **2023**, *17*, 1–20. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.