LABELANCE

A Decentralized Crowd-sourced Data Labeling Platform

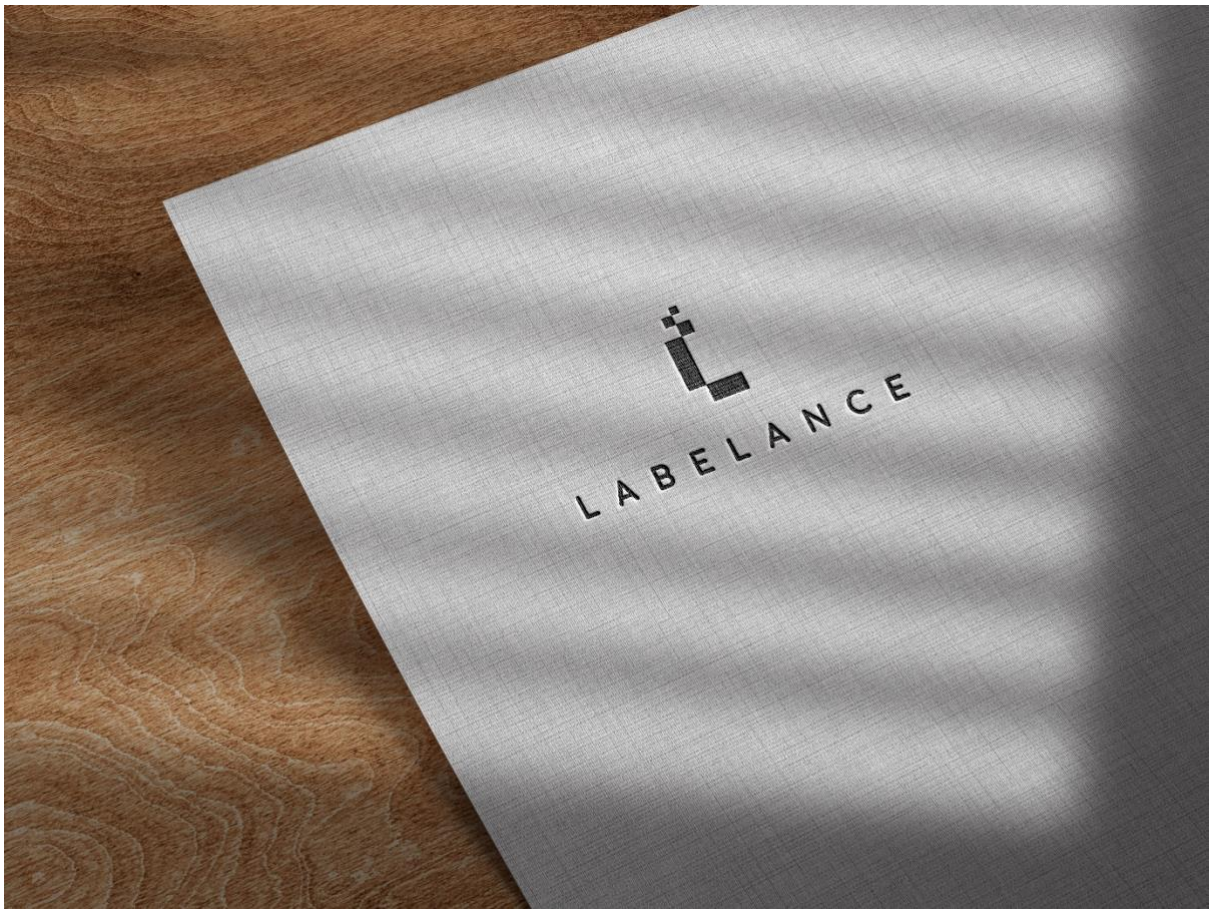

Table of Contents

1. Introduction **2**
2. What Is Data? **2**
3. What Is Data Labeling? **2**
3a. Text Labeling **2**
3b. Image Labeling **2**
4. Data Labeling in Data Science **3**
5. Labelance Ecosystem **3**
6. Data Labeling Quality on Labelance **4**
6a. Eliminating Spammers, Raykar & Yu Model **2**
7. Success Rates and Accuracy **3**
8. Web Application **3**
9. Gamification on Labelance: Magic Match **3**
11. Roadmap **3**
12. Use Cases **4**
15. Appendix **3**
16. Legal Disclaimer **3**

1. What is Data?

According to Oxford Dictionary, data means "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer".[1] Data comprises a wide array of useful information bits that can be used in decision-making, insight derivation, or making inferences. Data is also the main input for machine learning and deep learning models. Today, through the mentioned developments, data is aggregated at a massive scale by many businesses and governments to inform their decision-making processes. Nowadays, big data and machine learning techniques enable and empower people to achieve more, ranging from language processing to developing smart robots to artificial intelligence. Technological paradigm has evolved into a more digitized framework, and data fulfils the duty what once was done by coal. Data is what keeps the gears churning for today's technology.

2. What is Data Labeling?

Having mentioned what importance data holds in life, it is also imperative to recognize that, data leads to information, if and only if it is organized meaningfully. What renders data '*meaningful'* is the format in which it is stored. If the model at hand cannot comprehend or read the data, it serves no useful purpose. Hence, data should be in a form that can be recognized by computing systems. However, not all, if not most, data are not in such usable form. A sizable portion of collected data requires further handling to put it in an interpretable format. This is where data labeling arises. Data, either as text or numbers or images etc., is not necessarily aware of what it contains. Data labeling is the process of informing the data regarding its contents. This information is conveyed to data through putting '*tags'* on it. The tagging or labeling process changes from one format to other, the main processes which involves Paxos Data shall be briefly mentioned in this section.

3. Data Labeling in Data Science

Research suggests that data scientists spend a whopping 80% of their time preprocessing data and only 20% on actually building machine learning models[2]. Since the most time consuming and boring part of the task for data scientists are cleaning and organizing data, there is no doubt that crowdsourcing for data labeling helps data scientist in the matter of data labeling and tagging. Consequently, most of the data scientist embrace crowdsourcing and beseech having those tedious and weary tasks executed in the quickest way by those distributed workforces.

[1] "DATA | Meaning In The Cambridge English Dictionary". 2020. *Dictionary.Cambridge.Org*. https://dictionary.cambridge.org/dictionary/english/data.

[2] https://hackernoon.com/crowdsourcing-data-labeling-for-machine-learning-projects-a-how-to-guide-cp6h32nd

The survey of about 80 data scientists was conducted for the second year in a row by CrowdFlower, which is a "data enrichment" platform for data scientists. Here are the the results of survey:

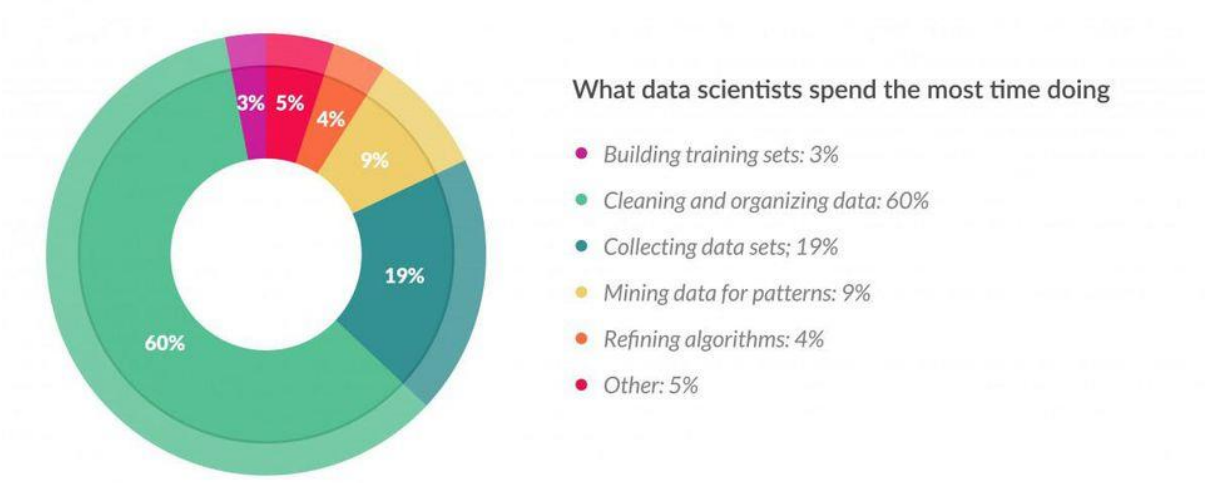

Figure 1. What data scientist spend the most time doing. From Forbes[3]

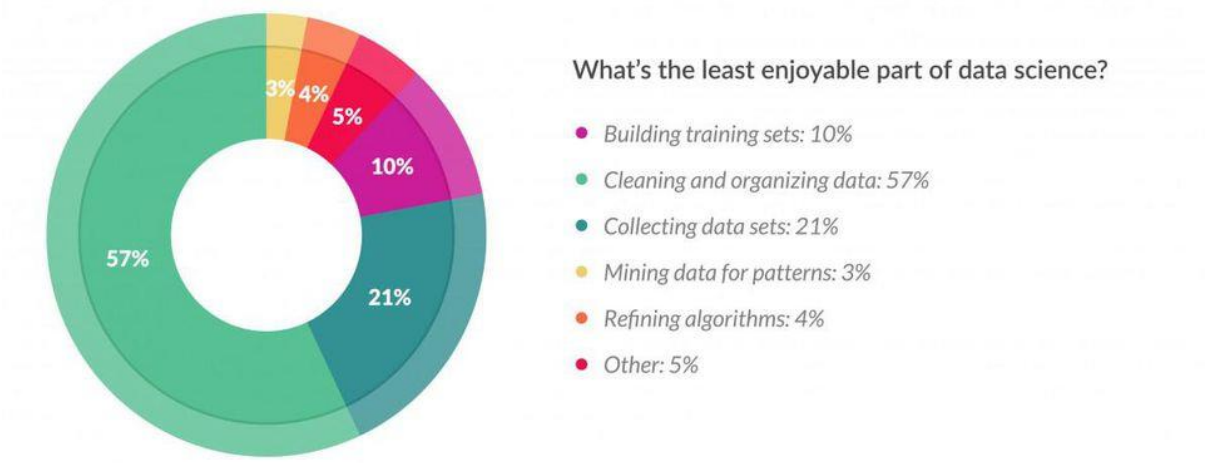

Figure 2. What's the least enjoyable part of data science? From Forbes[4]

According to above figures, data scientists spend 60% of their time on cleaning and organizing data and 76% of data scientists view data preparation as the least enjoyable part of their work.

[3] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?ref=hackernoon.com&sh=1a2313536f63

[4] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?ref=hackernoon.com&sh=1a2313536f63

4. Labelance Ecosystem

Labelance is a technology company providing crowdsourced data labeling solutions. Labelance platform offers solving data labeling needs of businesses affordably and create value for all parties involved. For the former we offer a fast and effective solution and for the latter we offer financial independence. By combining cutting edge technology with gig economy, Labelance empowers communities and smoothest the business process.

As Labelance, we are providing following data labeling solutions:

- Image Classification
- Bounding Boxes
- Text Input
- Text Segmentation
- Text Sentiment Analysis

6.1) Image Labeling on Labelance

Image labeling differs from its text counterpart in the sense of required information. Most statistical or computational models are not equipped to skim an image and determine what it is about. Even the most advanced machine learning models can only perceive the objects they are programmed for. Hence, it strikes as an important factor to know what an image is about or what it includes. For instance, an image of a full-fledged home and that of a run-down cottage may confuse some models. Even if they can discern a home from a cottage, they may not be able to distinctly recognize the attributes of the image of the house. It may not know whether the house has a direct view of the sea, how many floors it has, etc. Therefore, an image contains a plethora of information, and it depends on the objective to consider which information is relevant. Yet, utilizing the relevant information requires models to know that information is there. As Labelance, we provide the following data labeling methods for image types:

- Classification
- Bounding Boxes

Labelance mobile app has user friendly and easy to understand interfaces and designs for the newbies. New labelers can also read our handy user manuals and watch our instructive videos on our website, www.labelence.com. We warmly welcome anyone consider himself/herself that he/she can contribute those labeling services. Hereby, we aim to gather a great number of Labelancers in a little while.

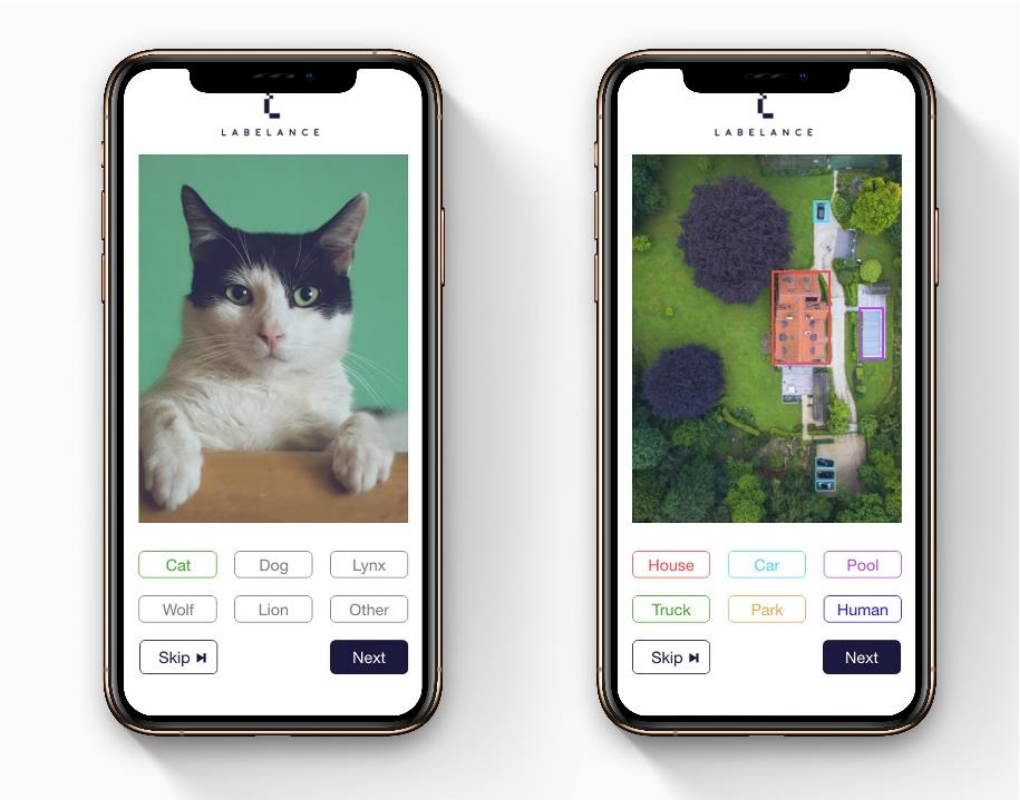

Image 1 - Image Classification and Bounding Boxes Operation on Labelance

6.2) Text Labeling on Labelance

As the name suggests, text labeling refers to labeling of data in text format. This can vary greatly according to the end purpose for the data. Thus, a text can be labeled regarding what emotion it conveys, its connotations, its sense of urgency, whether it is an abstraction or not etc. In brief, any classification of a text with respect to its qualitative aspects can be considered along the lines of text labeling. As Labelance, we provide the following data labeling methods for text types:

- Segmentation
- User Input
- Sentiment Analysis

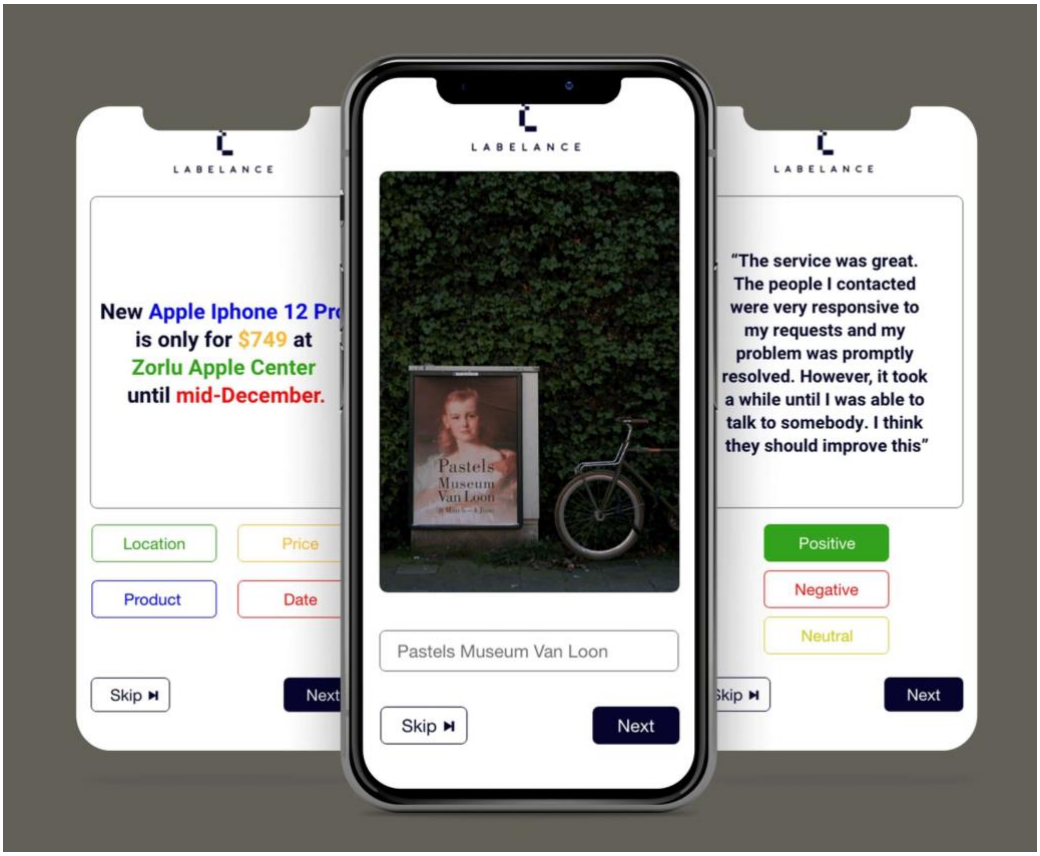

Image 2 - Segmentation, Input and Sentiment Analysis for Text Data

5. Data Labeling Quality on Labelance

A major drawback of most crowdsourcing services is that we do not have control over the quality of the annotators. The annotators usually come from a diverse pool including genuine experts, novices, biased annotators, malicious annotators, and spammers.[5]

7.1) Eliminating Spammers

In this chapter we are going to explain two methodologies for eliminating spammers. In this manner, spammer means a low-quality labeler who assigns random labels. Spammers can significantly increase the cost of acquiring annotations and at the same time decrease the accuracy of the final consensus labels. In this chapter we will explain Raykar and Yu's "Detecting and Eliminating Spammers Mechanisms[6]".

[5] https://dl.acm.org/doi/pdf/10.5555/2188385.2188401
[6] https://dl.acm.org/doi/pdf/10.5555/2188385.2188401

7.2) Annotator Model, Raykar & Yu

In this chapter we are going to reference Raykar and Yu's annotator model and, the article "Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks[7]".

Let $y_i^j \in \{0,1\}$ be the label assigned to ith instance by the jth annotator and let y_i be the actual (unobserved) label. Now, we model the accuracy of the annotator separately on the positive and the negative examples. If the true label is one, the *sensitivity* (true positive rate) for the jth annotator is defined as the probability that the annotator labels it as one.

$$\alpha^j \coloneqq Pr[y_i^j = 1 | y_i = 1].$$

On the other hand, if the true label is zero, the *specificity* (1 – false positive rate) is defined as the probability that the annotator labels it as zero.

$$\beta^j \coloneqq Pr[y_i^j = 0 | y_i = 0].$$

With this model we have implicitly assumed that $\alpha^j \; and \; \beta^j$ do not depend on the instance.

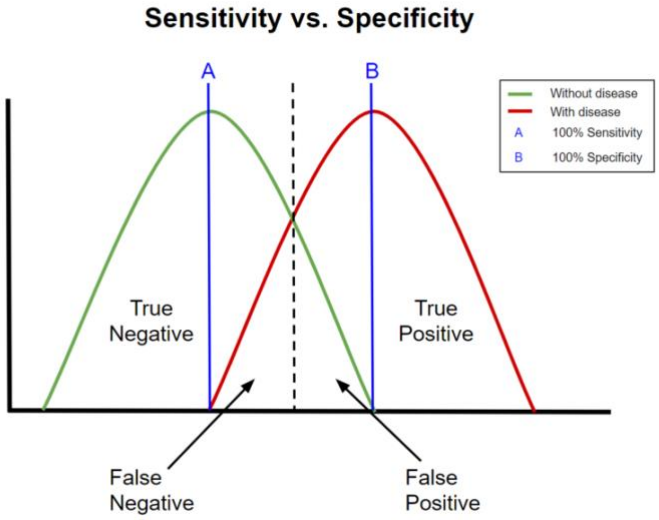

Figure 3- Illustration of Sensitivity and Specificity

7.3) Detecting Spammers and Ranking System, Raykar & Yu

Intuitively, a spammer assigns labels randomly, maybe because the annotator does not understand the labeling criteria, does not look at the instances when labeling, or maybe a bot pretending to be a human annotator. More precisely an annotator is a spammer if the probability of observed label y_i^j being one given the true label y_i is independent of the true label, that is

$$Pr[y_i^j = 1 | y_i] = Pr[y_i^j = 1] \qquad (1)$$

[7] Raykar and Yu, 2012

This means that the annotator is assigning labels randomly by flipping a coin with bias $Pr[y_i^j = 1]$ without looking at the data. Equivalently, (1) can be written as

$$Pr[y_i^j = 1|y_i = 1] = Pr[y_i^j = 1|y_i = 0],$$

$$\alpha^j = 1 - \beta^j \tag{2}$$

Hence in the context of the annotator model defined in previous section, a spammer is an annotator for whom.

$$\alpha^j + \beta^j - 1 = 0$$

This corresponds to the diagonal line on the Receiver Operating Characteristics (ROC) plot (see figure 3). If $\alpha^j + \beta^j - 1 < 0$ then the annotator lies below the diagonal line and is a malicious annotator who flips the labels. Hence, we define spammer score for an annotator as

$$S^j = (\alpha^j + \beta^j - 1)^2$$

An annotator is a spammer if S^j is close to zero. Good annotators have $S^j > 0$ while perfect annotators have $S^j = 1$.

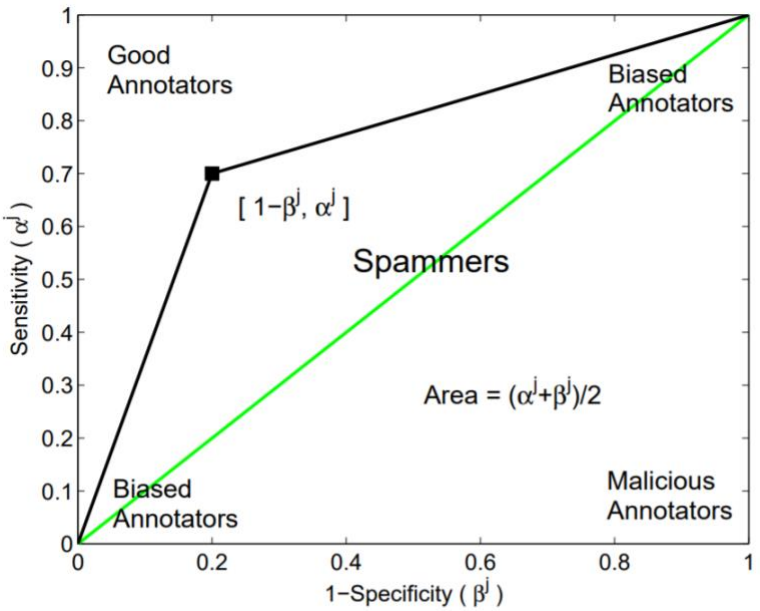

Figure 4 : For binary labels each annotator is modeled by his/her sensitivity and specificity. A spammer lies on the diagonal line on this ROC plot.

6. Success Rates and Accuracy

With the purpose of generating more confidential consensus structure, we periodically check our Labelancers' performances. Basic tests include accuracy calculations which is being applied for the eligible users after labeling certain amount of data. For security reasons, we will call this certain number as N. We also call total number of data labeled up to now by user as M and top score success rate as T (see next paragraph). The following formula will be applied after user labels N data whether the labels are true or false.

while M % N == 0:

    if data labeled > N

        if $S^j < 0.9$:

            flag user as "RED"

            qualityTest()

        else if $0.9 \leq S^j < 0.95$:

            flag user as "YELLOW"

        else if $0.95 \leq S^j$:

            flag user as "GREEN"

            if S^j > T

                rewardUser()

Users can check their success rates for each category via their profile pages. %96 and above means that labeler has a green flag, and between 90 and 95% user have yellow flag and below %90 shows a red flag. While green flag shows everything is good in countenance of labeler, the red flag has the contrary meaning of that.

Image 3 - Success Rates on Labelance

Labelance has determined lower limit of success rates and lower limit of top scores for each category. Labelers, whose success rates are below corresponding lower limits, will be put into a test which will not be indistinguishable for L. On the other hand, labelers, whose success rates are above lower limit of top scores will be rewarded with some amount of LBL token.

All these variances, lower rate of success rates, lower limit of top scores and rewarding prizes, are changeable due to the certain calculations.

For security reasons and better results, we are not going to announce those lower limits of success rates, lower limits of top scores and calculation methods anywhere.

7. Web Application

On Labelance web application, Labelancers can see their earnings, profile information and success rates. Additively, newcomers can take advantage of our handy user manuals for learning our data labeling tools and practice on some trial tests on the home page.

Labelancers can claim their earned Labelance Tokens and send to their wallets by simply connecting their wallets to our web application. In the 2021 Q1, we are going to launch Labelance liquidity pools and Swapping tool. Hereby, Labelancers can convert their Labelance tokens to major coins with few clicks. By 2021 Q2, we are going to present Locked and Flexible Staking facilities. So, Labelancers can benefit from high annualized earnings by staking services.

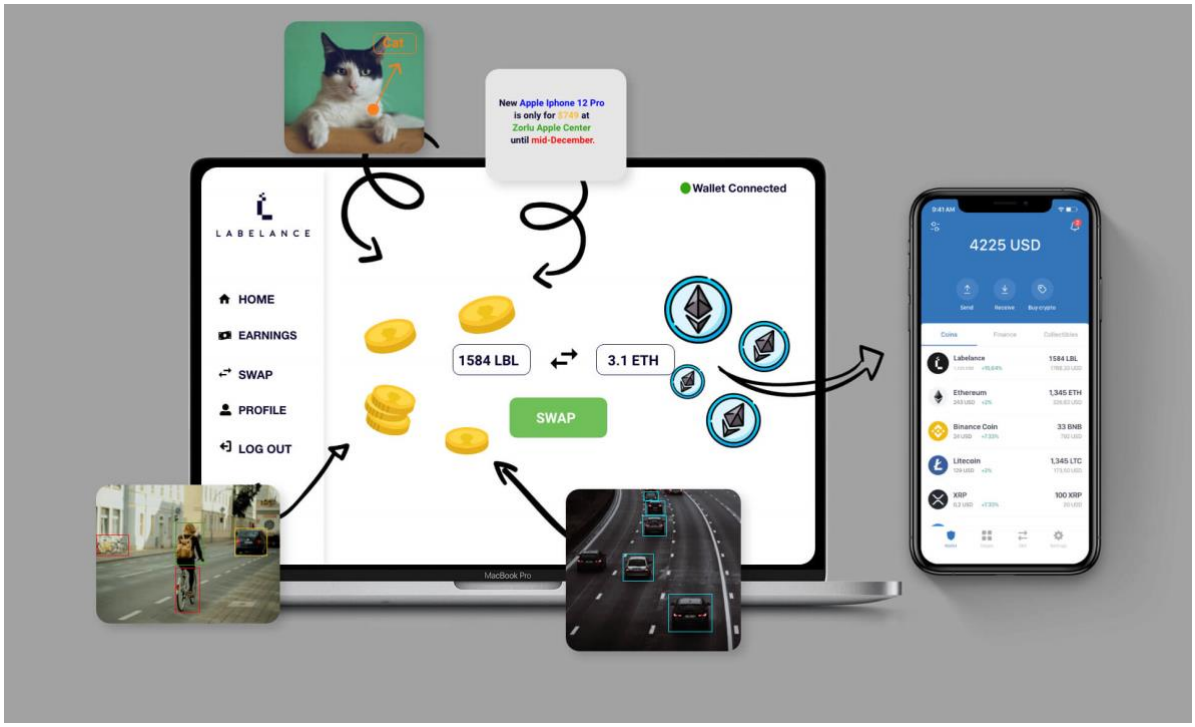

Image 4 – Labelance Web Application

8. Gamification on Labelance: Magic Match

Our goal is making Labelance efficient and sustainable. We believe that “the reluctant work” is not efficient and sustainable in all manners. Due to this, creating an entertaining ecosystem makes tasks easier to handle and the alacrity whole lot more. “Multi-play” and “matching” are the key words which upgrade the meaning of both competition and pleasure. Considering all these causative concepts, we created a game on Labelance: Magic Match!

Magic Match is a simple concept & one task in one step matching game. Unlike other games, the competition is not between the two sides of the game in Magic Match. There is a win-win situation in small scale which allows both sides to earn at the same time. In large scale, the competition emerges between all players because the top players will be granted varying awards.

The game’s structure is easy to understand and play. The matching players will see the same images on their screens, and they basically write the “objects’ names”. In an example scenario, user A and user B had matched, and an image was screened to both users simultaneously. Afterwards, users “randomly” type the objects’ names in the image that they see. Our matching engines use “synonyms dictionary”, “auto-correction” and various methods such as “the most commonly typed words together by mutual users” to increase true matching percentage. Let’s assume that there are four objects in the image stated above: mirror, chandelier, carpet and fridge. User 1 and user 2 begin to type as below:

User 1: “lamp”, “fridge”, “mirror”, “door”, “flower”

User 2: “mirror”, “chandelier”, “refrigerator”, “carpet”

In this situation, the same input typed by both users which is “mirror”, synonym words which are “fridge and refrigerator”, similar words which are “lamp and chandelier” will be matched by our matching engines and both users get 3 points. The other inputs typed by users which are “door, flower and carpet” will not be matched and users will not get any points.

Ultimately, the image matching operation will be done by various users and could be done in multiple times to increase accuracy rate.