Automatable quality assurance model for task solving within Crowdsourcing networks.

Mikkel Ole Rømer
Technical University
of Denmark
M.Sc. DTU Compute
S113408

ABSTRACT

Crowdsourcing provides a platform for outsourcing of tasks of computational hard nature to a great mob of human workers. A straightforward use case of this platform is labelling of data. However, as the tasks are distributed to a varying independent set of workers, the skill-set and knowledge of the working crowd will change from task to task. This drastically increases the complexity of assuring high quality of the task completion, while also making the cost per task somewhat incalculable.

This study seeks to address this varying quality phenomena within crowd-sourcing. In specific, this research will try to formulate a model optimised for consistent task quality while being cost effective. Moreover, the model must be automatable. The platform subject to this study will is the Amazon Mechanical Turk crowd-sourcing platform.

The paper proposes a straight forward 'get-focused' approach, which features a preliminary task aiming to force the worker to inspect the task at hand in detail. Through A-B split test of 3444 investigations, the approach described a quality improvement of up to 20 percent.

Source code for this paper available at:
https://github.com/mikkelscykel/turk/

Keywords

Amazon Mechanical Turk, Quality Control, Modelling, Imaging, Data Labelling

1. PREFACE

This paper was prepared at the Technical University of Denmark, DTU, as satisfaction of the requirements of 10 ECTS at the Digital Media Engineering study programme, summer 2016. The article is intended as a standalone paper, and no prior knowledge is required.

The Technical University of Denmark
2016

September 11th, 2016

2. INTRODUCTION

Machine learning and statistical prediction has shown immense potential. Each day new applications are revealing, utilising technical principles within this domain, in order to eg. imitate clever behaviour within software products. However, machine learning algorithms are a complex technology involving complex data, often with the need of clear training labels. Leaving the inspector with a potentially huge task of data-gathering. Multiple parameters must be considered when collecting data, among these are **Reliability** and **Velocity**. It is absolutely crucial, that the inspector gathers reliable data to provide qualified predictions. Moreover, the collection phase should be minimal and cost effective. Obviously face to face interviews are both expensive and time consuming. In fact, any approach in which a human inspector is involved in every entry collection is considered heavy, yielding the need of automated processes.

The Amazon Mechanical Turk platform (MTurk) is staging an online crowd-sourcing marketplace, in which the requester (the inspector) is capable of uploading Human Intelligence Tasks (HIT). A HIT is any sort of task, that the requester wish to see solved. A HIT is worth a small amount of money determined by the requester. When a HIT is submitted, it will be available to a large crowd of human workers. The worker, on the other hand, earns the credited money upon completion. In addition, the inspector can issue a bonus rewarding particularly good workers. The platform allows for, among others, fast data collection at a relatively low price usually ranging from 0.01$ to 0.20$. Practically, the platform provides a set of questionnaire templates; however, additionally it features an API and a IFrame interface, which allows for submission of basically any type of HIT.

When turning the investigation towards crowd-sourcing, issues regarding the data reliability reveals. Consider eg. the existence of spammers among the working crowd. That is, a worker who accepts HITs and submits random answers to collect the minimum HIT reward. As demonstrated [6, 8, 7, 3] the existence of spammers are a real threat towards data reliability, thus it must be addressed with high focus.

In addition, by the nature of crowd-sourcing tasks are distributed to a huge set of different workers. Workers with different backgrounds and skill sets. Clearly this obscures the precision of the data, when the task is subjective with no definite answer. An example could be rating of media quality. In this particular study, the workers are asked to rate a image containing food.

Though-out this paper, the MTurk platform is subject to

investigation. What happens when a set of random workers are asked to complete tasks for basically no money, and for a inspector sitting maybe thousands of kilometres away? What quality level can be expected? These are the questions, which will ground the basis of this investigation.

2.1 Related work

Nihar Bhadresh Shah and Dengyong Zhou [6] pioneered effective crowd-sourcing utility in 2014, with their specialised algorithm for optimising expenses while enhancing the outcome quality. In general, the paper proposes a *double or nothing* approach. The approach relies on a bonus scheme awarding the honest worker, whilst awarding minimal payment to spammers. Technically they rely on a number of hidden *gold* data points, which have been pre-labelled with their true label, also discussed by John Le et al. [4]. In addition they provide the worker with the possibility of skipping the question, if the he/she is not confident about the answer. The flow executes by providing a set of questions, for the worker to solve. The game then unfolds by introducing a number of gold entries among the questions. If a worker labels the gold entry with a false label, the bonus is reset. However, if the worker manages to label the entry with its true label, the workers bonus is doubled. The procedure relies upon the spammer submitting at random, thus the spammer will eventually hit a correct answer, however as the set of gold entries are increased the possibility of all randomly submitted labels being correct drastically decreases. Thus the paper strives to clean the input upon collection, instead of post processing the data. For further discussion on optimisation of crowd-sourcing data, please refer Shah et. al [6, 7], Grady et. al. [3].

Further study was done by Tian et al. [8], introducing more advanced machine learning to the domain, in particular they introduce an Bayesian Max-margin inference model, in order to enhance the quality of the crowd-labelled data.

Finally this study seeks to focus the workers attention prior to the actual investigation. Object based attention, selective attention, allows for an object to access our memory, Reisberg [5], this requires the brain to investigate the object and build associations. Focus priming principles will be used to enhance task solving quality.

2.2 Contributions

This study builds upon the *double or nothing* principle, proposed by Shah et al. [6]. However, as they only considered data with a definite answer, this study expands the principle towards more subjective data. That is, by utilising the knowledge gained from the definite answers as an indicator for the trust-value of the indefinite ones. Moreover, the investigations by Shah et al. relies heavily on simulations, whereas this paper supports Shah et al. by applying them to real world examples. In addition, this study provides a fully automatable approach, by utilising the Amazon API [2], suitable for real life applications where data-labelling is utilised eg. specialised search engines

Finally this paper proposes and investigates a 'get-focused' methodology, meaning that the users focus is primed towards the task, prior to the actual investigation, Followed by the discussion the contributions can be summarised as:

In specific this paper will

- Provide a automatable quality assessment algorithm enhancing the reliability of the data.
- Demostrate the *double of nothing* algorithm proposed by Shah et al. In the context of more subjective objectives.
- Investigating the attention priming 'get-focused' questionnaire approach.

3. ANALYSIS

This paper builds upon the observation, that each individual differs in both background, skills, intelligence etc. Thus naturally when crowd-sourcing work towards random individuals, whose only common denominator is an Internet connection, the results of that work will be of varying quality, consider example 1.

The goal is to provide a scoring mechanism, capable of identifying and priming the inputs of trusted high quality workers. Besides, identifying workers with a high trust score, it makes sense to maximise the set of trustworthy workers. The latter is true, since time and money are limited resources. Thus, it is of high importance to maximise data per price unit.

Though-out the following section, this paper will introduce the data subject to this study, explain the labels which it seeks to identify. This section aims at discussing how and why the solutions of this paper was chosen.

3.1 Disclaimer

This project does not investigate actual statistical predictions, nor does it experiment with any machine learning algorithms. The focus is to investigate trust worthy and precise data-collection from crowd-sourcing environments. Furthermore, a lot of different data measurements could be collected programmatically, examples of such are resolution, colour, contrast, no. objects in the image, background noise etc. These are all very meaning-full parameters when discussing imaging. However, likewise these are not correlated during this study.

3.2 The Data

As indicated this paper investigates data collection of more subjective nature. Thus the data subject to this paper is images. In specific, images of food. Table 1 provide basic information upon the data-set. The collection of data has been done in two ways.

The first an biggest set (A) is a set of photos submitted by random amateurs during a 2 week survey on Facebook. Specifically the participant was asked to take and submit pictures of food, which they would eat during the day. The second set, contains photos of taken by restaurant affiliates, that is people associated with restaurants and/or restaurant marketing. The two set contains only images of food, however their appealingness might differ. Generally the data-set varies both in size, resolution, orientation. An initial observation relates to the difference in quality related towards the images from the A tend to generally appear in lower quality, than that of the images from collection B. The difference is suspected, as the images from B is taken purely by people biased towards providing a great picture, as it represents their business, whilst the photographers of image set A is purely amateurs submitting images of what they eat to a social media. The reason behind having two data sets is: Firstly to improve the quality difference of the entire collection. Secondly to gather as many pictures as possible. An

Figure 1: Image 159 of the data-set A.

example image entry can be seen in figure 1.

Name	Number	Origin	Content
A	229	Photos taken by Amateurs	Food
B	84	Photos taken by Restaurant affiliates	Food

Table 1: The Data of this paper consists of 313 images, set A is a collection of images taken by amateur photographers. B is a set of images provided by restaurant associates.

As already bespoken, this paper deals with quality assurance of task solving within Crowd-sourcing networks. Furthermore, the task of attention is labelling of photos. Specifically, the intention is to ensure that the labels provided is trustworthy. That is not subject to falsified data. From section 2.1, this paper briefly described the principle of the *Double or Nothing* algorithm proposed by N. B. Shah et al. [6]. The algorithm, is specialised at identifying spammers, thus this paper draws on their research as spam detection. The very algorithm is dependent upon the existence of gold entries in the data set. That is, data points to which the true labels are known. For this reason, data-set A and B must include clear gold data entities. What is more, the investigation must provide the possibility for the user to indicate a neutral answer. The reason is, that if the worker is forced to click a definite answer to a question he/she honestly does not have the answer to, the analysis risk to label him as a spammer.

For this reason, this research introduces the category $Course \in \{Main\ Course, Dessert, Do\ not\ know\}$. The objective is to indicate if the dish presented is a main course, a dessert, or if the worker cannot tell. Obviously this category could introduce edge cases, where identifying a dish as the main course or as a dessert, might not be that straight forward. Thus the set of gold images must be chosen carefully, in order to eliminate the pitfall.

The main outcome by this investigation is to control the quality of a subjective label. That is, ensuring the highest possible confidence in a label with no definitive answer. Hence the paper introduces the interval $Appealing \in \{1, 2, 3, 4, 5\}$. The measurement is an indicator of how worker perceive the image, thus the measure is subjective in nature. Precisely, the worker is instructed to rate the image in terms of its quality and attractiveness. It is noticed, that while worker w_1 might perceive the quality as 3 another worker might perceive the quality as 5. The reason is the varying skill-set and backgrounds of the workers, consider example 1.

Example 1. Imagine two workers inspecting the same image. Worker one is a professional photographer, while worker two has no prior experience with photography. Obviously worker one is going to look at the image in a different way than worker two. Hence their results are prune to differ, when asked to inspect the quality of the photo.

Example 1 is illustrating the diversity of quality levels present in workers relating a specific task. In this case worker one as a higher quality level than worker two, providing the investigator with a much more professional answer.

Name	Description
Appealing	How appealing does the subject find the image: Scale from 1 to 5.
Course	Identify the correct category: Main course Dessert Do not know
Ingredients	The subject is asked to list the main ingredients of the dish

Table 2: The labels investigated by this paper.

Table 2 summarises the labels subject to investigation in this study. As seen a final category label called *Ingredients* is listed. The goal for the worker is to identify some of the ingredients which the dish is composed of. The purpose is to focus the worker towards the image, ensuring that the worker has properly inspected the image. That is, priming the worker towards a more genuine answer.

3.3 Minimising spammers with gold entries

This paper uses the Double or nothing principle, Shah et al. [6], as spam detector. However, before continuing the properties of a spammer must be defined.

Throughout this analysis, a spammer is regarded as a worker submitting random data to the questionnaire. This behaviour will obfuscate the available knowledge from the data, hence it must be detected and removed. Statistically this paper regards the inputs of a spammer as completely random. That is, in a binary investigation, the spammer will hit the correct question with a probability of $p = 0.5$. Likewise in a question with 3 possible outcomes, the spammer will be correct with a probability of $p = \frac{1}{3}$.

Shah et al. [6] proposes to introduce a set of gold data entries, in order to detect this behaviour. That is, a set of data entries, to which the true labels are known. However, this paper strives to gather information upon data of subjective nature, making true labelling significantly hard.

This could be comprehended over time by using a max-margin majority voting. Another way, is to introduce another deterministic label to the set. In reality this approach is seen all the time; eg. when users are asked to fill in a captcha upon submission. The functionality is effective and requires little effort in relation to majority voting. What is more, when introducing a democratic solution scenarios could arise, in which the majority is wrong. A situation like this could be when more spammers than real workers attend a task. This could have the devastating effect, of falsely labelling the wrong data as the true label. Since the ratio between honest workers and spammers i yet unknown, the captcha solution is chosen.

As bespoken, even the spammer will eventually hit the correct label, even with rather high probability. Obviously, one gold question per HIT is infeasible. The number of gold questions must be picked with caution. On one hand picking too few will allow potential spammers being classified as genuine workers, whilst picking too many drastically lowers the amount of new data collected during a HIT, hence increasing both expenses and time. Obviously this mechanism only works, if the gold questions are unambiguously and easy to understand. Binary questions are feasible, since it only has two outcomes thus minimises ambiguity and edge cases. However, as shown spammers are expected to hit the correct label with $p = 0.50$ chance. This analysis seeks to rule out spammers with a high significance that is $p < 0.05$. Thus the choice of gold questions needed is at least 5. Given true randomness among the spammers, spammers are expected to slip through with a probability of 0.0313 according to equation 1

$$p = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}^5 \simeq 0.0313 \quad (1)$$

Provided 5 gold questions per HIT, the total questionnaire grows. Acknowledging the price per data ration previously mentioned, the inspector should get more data out of the HIT than invested. As a result the total sample size of 20 is chosen. That is, 15 unlabelled or questionable data points along with 5 gold entries. Yielding a 15 to 5 ratio.

3.4 Addressing attention with a 'get-focused' approach

Focusing the workers object based attention prior to the investigation, would imply the worker to consider the image at hand before labelling. One way which has shown to enhance the attention, is to apply some level of interaction. Specifically, this study strives to obtain knowledge upon the image appealingness and quality. Thus, it is important that the worker inspects the image and understands its setting and contents. Implying the interaction should regard the detail level of the image. In order, to minimise development time and avoid distraction, the interaction must be simple and fast.

With the above parameters in mind, this paper proposes a free text categorisation. That is, provided a image of food, the worker is asked to list the main ingredients visible from the picture. This interaction, forces the worker to strive into the image components, what sort of dish is this, is it too blurry to even identify the ingredients, does the ingredients assemble fresh and healthy food and so on.

The 'get-focused' approach does not require the user to spend x amount of time. As the user is invited to proceed as soon as he/she feels that enough ingredients is listed. Reason being, that the focus could be lost if the user feels annoyed or stalled.

3.5 The Experimental Setup

This study was conducted using the Amazon Mechanical Turk Platform [1]. The platform has two main users; the requester which is the one providing the tasks. And the worker which is the user solving the tasks. The worker can be any person with an Internet connection, making the skill-set very scattered. The setup created in this study, is a Django website with a mysql database. The website hosts the questionnaire, which is presented to the user.

Besides the data which is acquired from the questionnaire, Amazon provides yet another valuable information, namely the identification of the user (the user id). This information, allows the requester to track the workers, thus providing the opportunity of collecting user insights upon their skill-level.

From figure 2 the overall flow of the investigation is illustrated. From the figure, it should be observed that the procedure distinguish users into two categories, defined below:

1. $\overline{U}$ Representing unknown and or low quality workers.
2. $\overline{U'}$ Representing known high quality workers.

Notice that, throughout this analysis, low scoring and new unknown users are not distinguished. This implementation choice is rooted in the very concept of managing a user score. In fact, the entire set $\overline{U} + \overline{U'}$ are stored and managed the same way. Thus the difference is managed by adjusting the threshold τ according to section 3.4. Further discussion of τ follows from the result section.

The investigation was conducted using the entire data set, table 1, that is a total of $i = 313$ images. In addition, each HIT features $t = 20$ different images, and a total of $n = 180$ HIT's were conducted. Providing that on average each image was presented 11 times, refer equation (2). Moreover, a total of set of $\overline{G} = 30$ gold images was selected prior to the investigation. The entire experiment setup is summarised by tabel 3.

$$\frac{t \cdot n}{i} = \frac{20 \cdot 180}{313} \simeq 11 \; views \quad (2)$$

Total data entries	313
Total gold entries	30
HIT size	20
Number of HIT's	180
Threshold	0.80

Table 3: The table summarises the experimental setup, executing a 180 HIT's.

4. RESULTS

There are two scenarios suitable for discussion, when dealing with crowd-sourced subjective data. As mentioned previously, obviously the inspector is likely to see data, where the subjects disagree upon its label. Figure 3 visualised this scenario. Approximately half of the 11 workers, who saw

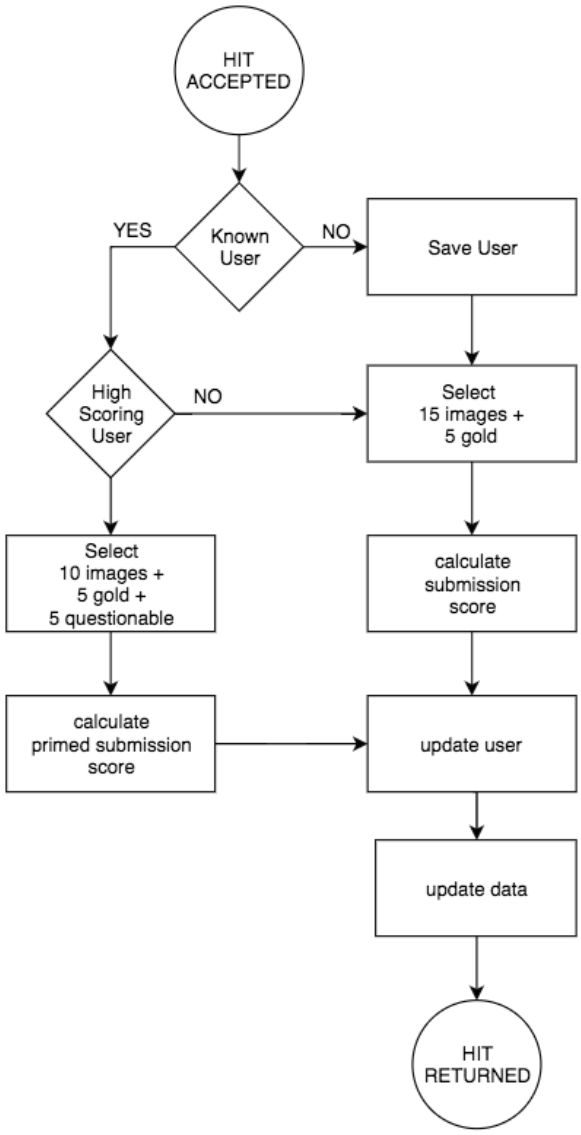

Figure 2: The images illustrates the overall test flow, notice that the result is primed when the user is known to be trustworthy.

this image, valued it with the very worst score (1). Contrary the other half seems to rate this image as the next best thing (4). The second scenario, involves the presence of the spammer. From figure 4 such a scenario is illustrated. An important observation is, that this investigation seeks to remove outliers, figure 4, while it does not try streamline peoples opinion, figure 3.

As mentioned previously, this investigation implemented a free text label called *ingredients*. The task is for the worker to list the main ingredients visible from the picture. It was argued that this label, would force the worker to inspect the image properly, thus priming the focus minimising the probability of incorrect labels. This however must to be tested, leading towards hypothesis (1).

***Hypothesis** 1.* If a worker is forced to inspect the task in details prior to labelling, the worker will perform with higher quality.

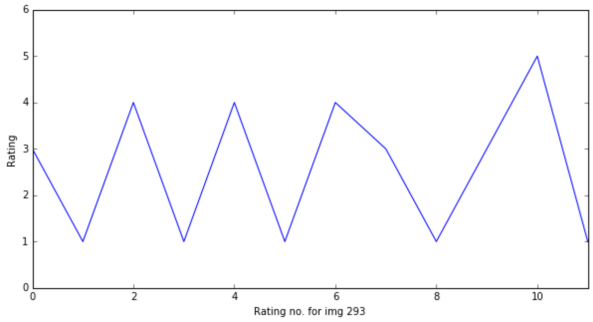

Figure 3: From the graph it is clearly seen, how the workers on this particular image disagree widely upon the image quality.

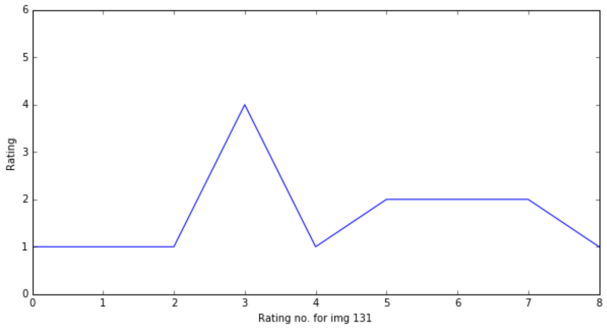

Figure 4: The graph visualises rating scores of image 131. In this particular case a spammer is responsible of disrupting the trend.

To investigate hypothesis (1) this paper makes use of the well known randomised A-B split test. The workers presented with test A, t_A, is asked to label the image before asked to identify the image components. The worker is allowed to change his/her answer of the initial labels at any time. Contrary the users of test B, t_B, is asked to identify the image components prior to the labelling process. Likewise, the user can modify any answer at any time until submission.

Figure 5 summarises the findings. From the figure the quality performance, calculated using gold data entries, of the workers are visualised. That is, the y-axis represents the cumulative percentage of workers performing at or above the score visible from the x-axis. The upper green line represents t_B, whilst the lower blue line represents t_A. Two observations should be made. Firstly, from inspection of figure 5 it is seen how well the distribution of t_A and t_B assembles each-other. In fact, the trend lines are almost completely symmetric. Secondly, it is readily observed that the distribution of t_B is shifted upwards with a factor between $\simeq 1.07$ and $\simeq 1.20$.

Specifically this illustrates how the number of workers present in each quality group grows close to linearly, figure 6, when the workers are forced to inspect the images in detail prior to the label phase. As an example, by inspecting figure 6, it is seen that the set of workers achieving $>= 0.85$ worker trust, that is workers who provides true-labels in at least 85% of the questions, are 20 percent bigger in set t_B than that of t_A, drastically enhancing the data quality, confirming hypothesis 1.

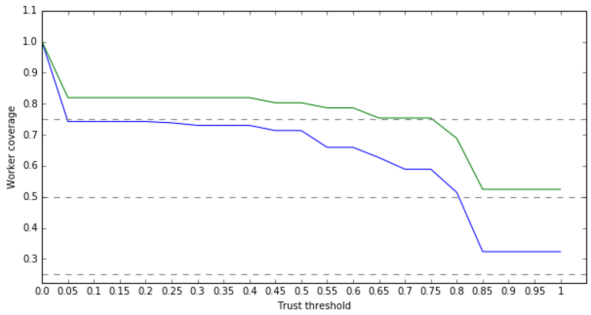

Figure 5: The plot illustrates the worker coverage distribution in relation to the worker-trust-threshold (WTT). The upper line represents data set t_A and t_B, lower and upper respectively. It is readily seen that the data-sets assembles similar distributions, however t_B is shifted upwards. Indicating higher quality.

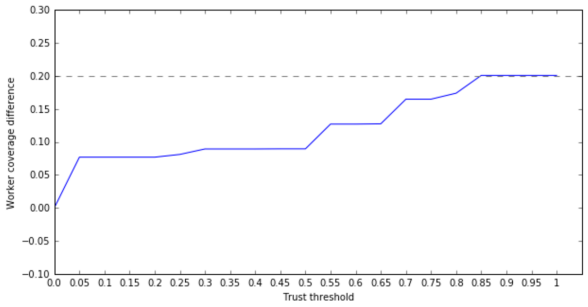

Figure 6: From figure 5 the two distributions t_A and t_B are plotted against each-other. This plot illustrates how the quality of the workers grows close to linear when forced to inspect the image prior to the label phase.

Likewise it makes sense to look at the rating distribution of t_A and t_B, since the underlying data are identical, and the trust-score much higher in the latter case. Figure 7 and 8 visualises the rating distribution of t_A and t_B respectively. An important observation is that the ratings of t_A reflects much higher positivism. That is, in general workers of t_A rates the images higher than the workers of t_B. From previous sections, table 1, it is seen that the vast majority of images present in this investigation origins in amateur photography, involving phone cameras, low personal involvement, no or little focus on setting etc. Thus, the size of low and medium scoring images are expected heavily. This expectation, is better represented by t_B than t_A. It must be noted however, that the true quality distribution is not known, thus the correctness of t_B over t_A is indiscreet. Thus the observation can be seen to back the assumption, however further investigation must be performed in this setting.

As discussed, the worker is likely to perform significantly better when forced to inspect the image prior to the questionnaire. However, what is the actual detail level of this free text exercise? Earlier this paper presented a image from the data set, that is image 159 seen in figure 1. It is of interest to access the quality of this field, as it has proven to enhance quality of other fields.

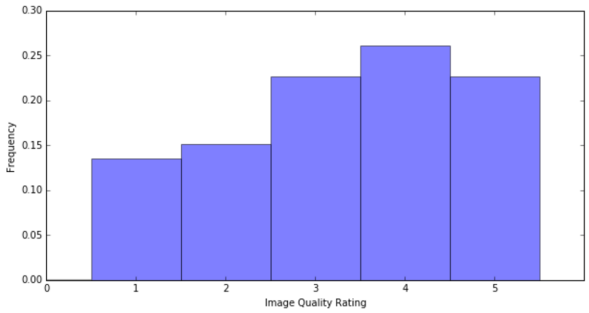

Figure 7: The rate distribution of set t_A. The distribution is skewed in positive direction, implying that workers are finding the majority of the data-set in high quality. Recall table 1 expecting the majority of the data-set being of amateur nature.

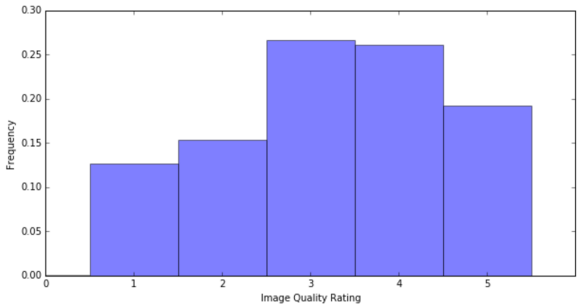

Figure 8: The rate distribution of set t_B. The distribution, like t_A, is skewed in positive direction. However, centred better around the mean score. Implying, higher level of criticism in set t_B than t_A.

From figure 9 the word frequency of all entries in this field, regarding image 159, are plotted. The entries have been stemmed, tokenized, and in some cases spell corrected. Image 159 is made from the following ingredients:

$$I \in \begin{cases} Pasta\ Penne \\ Rucola \\ Cucumber \\ Cherry\ tomatoes \\ Cheese \\ Red\ pepper \\ Red\ onions \\ Walnuts \end{cases} \tag{3}$$

Remarkably all ingredients listed in (3) are found within the 11 most identified components, disregarding the colour artifacts red and green. In fact, only one incorrect token reached the top, namely noodles. This ground basis for training of highly clever machine-learning algorithms. Since the data provides real names and not only pasta with salad. Identifying components using machine vision, allows for highly intellectually search engines. Since it provides the ground of guessing kitchen origin, dish name, etc. A reasonable goal for improving the results of this free data, would be to improve the questionnaire when initial data has been obtained. For instance, changing into a quiz approach asking does the image contain ingredient x. When suitable data basis is ob-

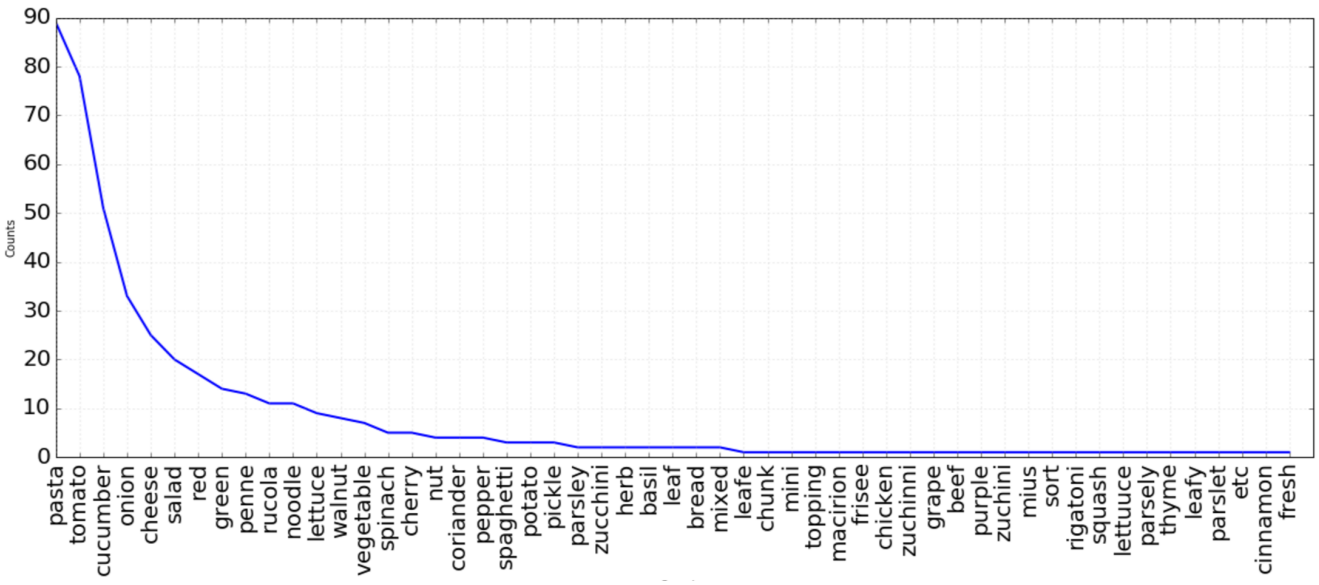

Figure 9: This figure assembles the word-frequency of the free-text field of image 159, visible from figure 1. The distribution is of all ingredients submitted by any worker on image 159 regardless of worker trust score. The only modification made a word stemmed tokenization, along with a spell corrector.

tained, it could develop into a more interactive approach. asking the user to mark a specific ingredient.

5. CONCLUSION

This study investigated an automatable process assuring quality when utilising crowdsourcing. The main problem grounded in two domains, namely the existence of spammers among the crowd, and the varying qualification levels among the workers, figure 4, 3 respectively. This study used the observations reported by Shah et al. to deal with spammers. In addition, this paper introduced a scoring mechanism, allowing the requester to identify trustworthy workers, and thus enhance the overall quality.

The latter case is, in particular, good when investigating tasks of subjective nature. Throughout, this paper the subjectivity was provided through a quality inspection of images, however the principle could easily be extended towards other tasks.

What is more this study, suggested and proved, how providing a preliminary task to the worker, with the goal of forcing the worker to inspect the task in detail, resulted in a quality improvement of up to 20 percent. In fact, in this case the preliminary task in itself generated highly sophisticated data, grounding basis of intelligent machine learning.

Finally, this paper suggested that the unfocused worker has a tendency of skewing ratings in a positive direction, a knowledge which could be subject to a future investigation. As it could allow for enhanced automated data cleaning. That is, if it is possible to measure and understand the very skewness. Additionally, this paper proposed a weighted trust scoring system, allowing the data to benefit from the professional worker, whilst learning bits bytes from the non-focused worker. In collaboration with Shah et al. this could not only maximise data precision, but minimise costs as well.

The analysis at hand is however preliminary. That is, only a small real life test sample was investigated. Hence further study should verify and refine the scoring mechanism. Moreover, the time of day, at which the analysis was investigated should be expanded, in order to normalise users across time zones. In addition, further investigation would drastically highen the change of observing the same worker, drawing even further benefit from the procedure.

A final remark leads towards the features which is needed for an automatable process to be feasible. An automatable process would often be needed in a real time environment. Which on the other hand means, that the Mechanical Turk platform should provide reasonable response time. That is the time it takes for an HIT to be completed. In this experiment 3444 images was processed in approximately 382 minutes, which is about 9 per minute. However, by inspecting the server log it is observed that many workers accept the HIT to solve it later. The HIT's submitted in this paper had a seven days time to live, hence it would be interesting to see if a speed up would be achieved by decreasing the time to live, forcing the worker to complete faster.

6. REFERENCES

[1] Amazon. Amazon mechanical turk. In *MTurk*. 2016.

[2] Amazon. Amazon sdk. In *Python 2.7*. 2016.

[3] C. Grady and M. Lease. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 172–179, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[4] V. H. John Le, Andy Edmonds and L. Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *In SIGIR 2010 workshop on crowdsourcing for search evaluation, pages 21–26*. 2010.

[5] D. Reisberg. *Cognition Exploring the Science of the Mind*. W. W. Norton and Company Limited, 2015.

[6] N. B. Shah and D. Zhou. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1–9. Curran Associates, Inc., 2015.

[7] N. B. Shah and D. Zhou. No oops, you won't do it again: Mechanisms for self-correction in crowdsourcing. In *Proceedings of The 33rd International Conference on Machine Learning, pp. 1–10*. 2016.

[8] T. TIAN and J. Zhu. Max-margin majority voting for learning from crowds. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1621–1629. Curran Associates, Inc., 2015.