NOT THE LAST VERSION

REMOVED
BLIND VERSION

***Abstract*—The abstract goes here.**

Keywords*—*crowdsourcing; workload

***** NOT THE LAST VERSION****
I will continue off-line...

I. Introduction

Please do not use items...

XXXXX

On the basis that Crowdsourcing is a growing market, different aspects in these platforms have been analyzed in the past years.
A 5 year study was conducted to understand crowd and improve the platform to make it more efficient for workers and requesters[1] In this research, there are 29 features analyzed trying to explain changes and conducts of the crowd during this years.
In order to improve this analysis and perform a better throughput and a complexity prediction workers were asked to execute some tasks and after that they had to fill a NASA TLX[2] questionnaire. With this results and the visual and content features that they have extracted from the different HITs they have made a model to approximate the complexity of the tasks. The idea was also to improve the platforms by helping the requesters to judge how complex a task is for the workers ind order to adapt it. [3]
The results show that for the prediction, the type of HIT is a central key-point. The predictor of [1] was used and improved adding the complexity of performing the Task.
After this the idea is to analyze some more and different features (visual and content) trying to find some new key-features. Another big difference is that in the actual study the workers only see the task, they don't perform it, the idea behind this is to analyze the perceived workload of a task by the crowd.
Performing Micro-Tasks in Crowdsourcing platforms like MTurk is gaining popularity in the last years. The huge amount of data that is created daily now a days needs to be preprocessed. As computer algorithms are not performing optimally for all tasks so businesses use Crowdsourcing platform to perform the so called Human Intelligence Tasks (HITs), such as video tagging, image description, text translation, surveys, etc In these platform requesters create HITs with small tasks and the crowd (workers) can apply to them in exchange to a small reward per HIT. Its a fast growing market, as much as data grows.
The process starts with the requester creating a HIT basing on the task and data that he wants to obtain. After this the HIT will be uploaded in the Crowdsourcing platform as a Batch of HITs with a given reward and requirements of the workers. In the last step the crowd perform and submit the tasks, the requester checks the results and he pays the workers.
The Throughput rate is defined in Crowdworking as: The rate of HITs that get completed between two successive observations. (HITs/min.) This is measurable. This rate is predicted analyzing the metadata of the HITs, such as: Title, Description, Keywords, Reward, Date, Allocated time, Batch size. The paper that was analyzed[1] explains that a filtering of the tasks in different categories was needed to improve the prediction method.
After filtering the Data, a correlation and a feature selection are needed. A Random Forest Regression, with variation of the window size, was used to finish the prediction model.
The analysis show that large batches are better to use for prediction of worker behavior and have more throughput. There are 2 important features: HIT_AVAILABLE and Age_Minutes. The HIT_AVAILABLE feature describes the number of tasks per batch and has a weight of 40%. The other important feature with a weight of 20% is Age_Minutes, here it is described how old a HIT is.
In order to improve the analysis of crowdsourcing the aspect of *Complexity* was analyzed.[2] For the prediction visual features, content feature and metadata were used.
Visual features were extracted from the website itself, number of photos, color of layer, layer type, size of photos, text groups, etc..
Also content features such as semantic richness, words count, links, topics, keywords, etc
The conclusion of this study reveals that the throughput prediction is improved by using the complexity factor.
Just check this, have make my best, Vic
Taking this facts the proposal of this paper is to model a predictor of the perceived workload. The focus is on how workers choose the tasks only from the information they have before they submit to it, what workload perception they have from it. It's an important approach to improve the design of the tasks and for understand how workers choose the tasks. On addition a prediction of the acceptance of the task was done. The goal is to predict if a job is going to be accepted by workers just having the information and the preview of it.

A. Workload

Workload ”represents the cost of accomplishing mission requirements for the human operator”[3]. Beside other factors workload influences the job choice decisions [ref]. Real and perceived workload differs. Which latter being influencing expected compensation. Fair balance between workload and compensation leads to increasing productivity and throughput rate [1].

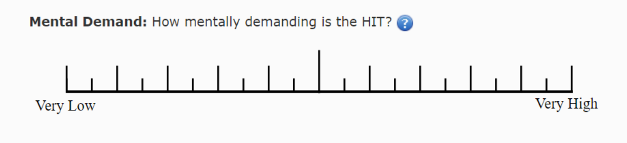
Previous studies show that badly designed jobs in crowdsourcing microtask leads to collect low quality responses [2]????. A growing body of literature has addressed ethics including fair payments in current crowdsourcing microtask practices [4], [5], [6], [7], [8], [9]. Being able to measure workload corresponding to a microtask is the first step towards fair crowdworking.

In the first conception of workload the key measure was the physical demand. In the actual world heavy work is attached to machines, and the studies analyze other types of workload, psycho-motor, perceptual, or communication workload [10]. Different techniques are used for the measurement of workload including but not limited to subjective measurements (e.g. [subjective workload dominance, Bedford, NASA-TLX])[11], Psycho-physiological measurements (e.g. [Heart Rate, Heart Rate Variability, Evoked Potentials])[11]and ... . For subjective assessment of workload, usually participants performing the target task and afterward asked to fill in a questionnaire containing measurements items. Subjective measures can be divided into unidimensional and multidimensional measures depending to the scale used for rating. Unidimensional scales are simple and fast but more sensitive than the multidimensional scale [12]. The multidimensional workload assessment scales are more complex and are used for diagnostic [12], e.g. SWAT [13] have three dimensions and NASA-TLX with six dimensions [3]. In this study the *Rating Scale Mental Effort* (unidimensional rating) and the *NASA TLX* (multidimensional-ratings) were employed.

1) NASA-task load index (NASA-TLX): It has been developed in 1986 [14] and is one of the most widely used multidimensional scales for measuring the workload [3]. It consists of following six subscales and hypothesis that some combination of them are likely to represent the workload experiences by participants [3].

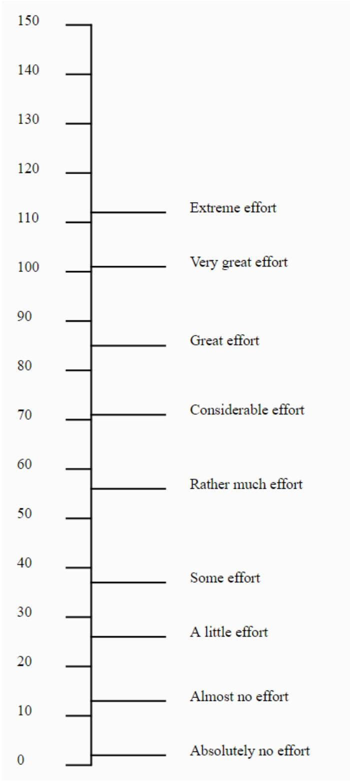
- Mental Demand: How much mental and perceptual activity was required?
- Physical Demand: How much physical demand was required? (physical work)
- Temporal Demand: How much Time pressure is perceived during the task.
- Effort: How hard was the task to perform?
- Performance: How do you perform the task? (fail, success)
- Frustration: How do you feel during the task? (stress, irritation, happy, etc..)

As individuals, with different skills, perceive influence of each dimension differently on their experience of workload, an individual wighting schema was created [14]. The weights are drived for each participant by using a set of simple pair comparison between six dimension. In each one participant should decide which dimension is more related to their personal definition of workload [3]. As a result final calculated workload is tailored to individual workload definition.



Mental Demand: How mentally demanding is the HIT?

2) Rating Scale Mental Effort (RSME): In the Netherlands, a unidimensional scale, RSME, was developed by Zijlstra (Zijlstra & Van Doorn, 1985, Zijlstra & Meijman, 1989, Zijlstra, 1993).
The ratings are placed in a scale from 0 to 150 over a continuous line. Every 10 there is a mark with statements related to invested effort, e.g., On the RSME scale the amount of invested effort into the task has to be indicated, and not the more abstract aspects of mental workload (e.g., mental demand, as is in the TLX). These properties make the RSME a good candidate for self-report workload measurement.



II. Method

A dataset containing samples from various HIT types were collected, and estimated workload for each HIT was assessed by crowdworkers. In addition to the workload related scales, workers assessed other aspect of each HIT including interestingness, difficulty, complexity and whether they accept to perform it given the specified reward. Next, relevant features were extracted. Different models for predicting 1) workload and 2) acceptance of HIT were created and evaluated. In the following each step is described in details.

A. *Dataset*

In the first step, 400 different HITs from MTurk were collected in November and December 2016 by four individuals. Main criteria for considering a HIT was to be able to estimate its workload from the preview page. Thus, surveys with external links and multi-page HITs were discarded. For each HIT, the screenshot, the complete HTML source code, and the available meta-data (e.g. title, description, rewards, time allocated, available HITs) were collected. Further review by a fifth person leads to 359 HITs approved to be included in the dataset. From them, 14 extensive HITs, which have repeated patterns of questions, were selected and particularly *manipulated* in terms of their length. Consequently for each of them three different instances were added to the dataset namely in the original length, 50% of the original length, and 25% of the original length. For manipulated set of HITs, it is expected that length of the HIT positively correlated with its estimated workload. The purpose of including the manipulated HITs in the dataset is to provide an instrument in order to later evaluate the reliability of the subjectively assessed workload. As result, the final dataset contains 401 HITs.

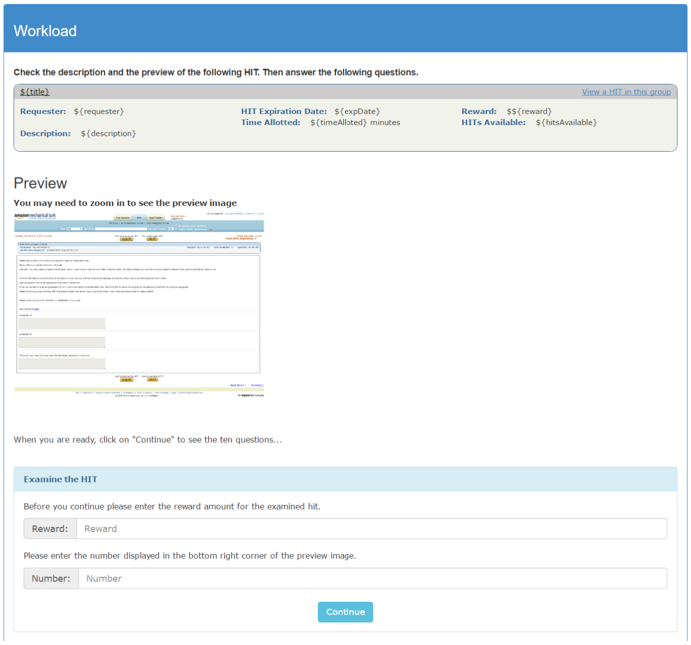
In the second step, all HITs in the dataset were categorized following the schema suggested by Gadiraju et al. [15]. A crowdsourcing study was conducted in MTurk. For each HIT from the dataset, meta-data and the screenshot were presented to three different workers and they were asked to select the type of HIT from a given list of selection (i.e. 6 categories proposed by Gadiraju et al. [15] and also ”Do not know”). Job was available for US workers, who has more than 500 approved HITs with the overall approval rate of 98 % or more. Same Criteria is used for the rest of crowdsourcing studies as well.

The job was rewarded by $ 0.02. Answers from all workers were accepted. For 161 HITs out of 401 all three workers agreed on the HIT type. The study was repeated for the remaining HIT and again three more workers determined the HIT Types. Finally, 56 HITs that workers did not agree (less than three votes) on their type were judge by an expert.

1) Introductory Job: The introductory job was a prerequisite for the main study and was available for 200 workers. It consist of two parts; first demographic and motivation questionnaires and then 15 pairwise comparison questions referring to the six subscales of NASA TLX [3]. The job was rewarded with $ 0.15 and finished within one day. Two trapping questions [16] were employed in the first part and two of the paired comparisons were repeated to check for reliability and consistency of results.

From the respondents, 12 were answered one of the trapping questions wrongly, and 31 had inconsistency in their choices (i.e. pairwise comparisons). Although all respondents were paid, 157 workers were qualified and invited to the next job.

2) Workload Assessment Job: After sending an email inviting all the 154 workers who were given access to the second job they could start rating the HITs that we collected . Of the 401 HITs five had an issue with showing the preview screenshot properly. To incentive the workers to rate as many HITs as possible a bonus payment of 0.05$ per rating was added to the 0.10$ per job if they performed more than 20 HITs.
In the job the information which was given to the workers was some of the metadata :Title, Requester, Description, HIT expiration date, Time alloted, Reward, HITs available and the preview screenshot:



Workload

Check the description and the preview of the following HIT. Then answer the following questions.

${title} View a HIT in this group

Requester: ${requester}
HIT Expiration Date: ${expDate}
Time Allotted: ${timeAlloted} minutes
Reward: $${reward}
HITs Available: ${hitsAvailable}
Description: ${description}

Preview

You may need to zoom in to see the preview image

When you are ready, click on "Continue" to see the ten questions...

Examine the HIT

Before you continue please enter the reward amount for the examined hit.

Reward: Reward

Please enter the number displayed in the bottom right corner of the preview image.

Number: Number

Figure 1. Screenshot of the Workload assessment job as seen by the worker

After entering the reward and a random number which we generated in the bottom right corner of the screenshot the workers could start rating the HIT. First they were asked a binary question whether they are able perform the job. To ensure that they are able to judge the job. After that 11 questions had to be answered about the job consisting of the six NASATLX dimensions and 5 additional ones:

- how interesting and enjoyable is the HIT (itself) for you?
- how often do you perform these kind of HITs?
- how difficult is the HIT for you to perform?
- how complex (too involving, complicated, confusing) is this HIT for you?
- how likely are you to work on the HIT?

Each of the question were answered on a 20 point graded scale on which the low indicator (very low, failure, not at all, never, easy, simple) was on the left and the high indicator (very high, perfect, extremely, often, hard, complex, very likely) was on the right. A trapping question was included were the rate of the HIT had to be entered and we asked the workers to estimate the time it would take them to perform the job. Finally every worker had to rate the overall effort for the HIT this was also determent on a graded scale this time with 16 point reaching from 0 to 150. Each of the HITs was rated 5 times 2001 answers were collected.

Analyzing the Data that was collected 2 observations were made:

- The Standart Deviation of the TLX Values in some cases was very high.
- NASA TLX and Overalleffort correlated high.

Probably some workers answered randomly. The first problem will be later in the section *Data Cleaning* solved. Taking the other observation in consideration the Label to predict was then Overalleffort. A new job was created to have better accuracy in the Mean and the Standart Deviation asking for

the Overalleffort.
TODO @Victor

B. Feature extraction

Three group of features are extracted and hypothesized to be useful for predicting estimated overall workload. *HTML based Features* are extracted by parsing corresponding HTML code and counting different HTML tags like *input*. *Visual Features* are extracted from the HIT preview and mostly focus on how dense is the text content of task. Third group of features are extracted based on Natural Language Processing (NLP) concepts referring to the XXXX. In the following each group of features are explained in details.

1) HTML based Features: TODO @ Philipp The HTML-based features were extracted by looking at the source code of the HITs. All of the feature can be assigned to one of two categories:

Syntax analysis:

In the syntax analysis the we try to obtain information about the complexity of the task based of the SGML (Standard Generalized Markup Language) elements of the code. Those SGML elements are marked by tags which give clues on the type of media which is used in the HIT or what kind of input is required to finish the task.

The <*img* >, <*audio* > or <*video* > tag can give some numerical information from the size of the file and the amount of files which are used. That indicates how much information the worker has to process before he or she can start the task. A slightly more ominous media type that frequently appears is the <*href* > tag which implements links. Here it is more difficult to make judgments on the workload because the code doesn't revel the content behind the link.

The <*select* > and <*input type =radio* > tags indicate multiple choice answers. Here fewer answers indicate a smaller workload due to less reading effort and an decision finding. Another common input among the HITs is the <*input type :text* > which lets the worker post a free text. Here is also the workload very difficult to judge without a semantic analysis because on the tag alone one cannot obtain how much text is required to finish the task.

But the tags do not only give information about the task itself but also on how its presented. We can identify the size ,style and color of the font. Whether it is bold or cursive. All of those can give valuable information on how difficult to grasp the task is for the individual worker.

Semantic Analysis:

In the semantic analysis the syntax is completely disregarded and only the text seen on the web page by the worker is sifted through for information. In other words the program is trying to derive the meaning of the task by the code. This is called natural language processing.

Parsing:

In natural language processing one of the major tasks is parsing the text. That means the symbols ,letters and punctuation marks, are put in a string and than analyzed grammatically and than put in a parse-tree. The parse tree is a structure in which the desired properties of the text can be deducted by the program.

This includes extracting the number of words ,amount of sentences ,sentence length and length of the by-sentence. The idea is if the worker has to read a lot about the task or it was explained in a complicated way that would increase the workload of the HIT.

More difficult deductions are linked to the vocabulary. So that there would be a better indicator on how easy the task is to comprehend. For this task a library of 'difficult' vocabulary is needed with which the words of the given task is checked for analogies or check for the number of unique word stems.

2) Visual Features: Visual features were extracted by processing the screenshot of the individual HITs (cf. Section XX). In the field of Image Processing, a common measure to extract meaningful image regions that stand out from other parts is to perform "saliency detection"[17]. Saliency at its core is just a collection of features of an image. These features include quantities like color gradient, and edge frequency in the image. Saliency itself is a measure of the important parts of the image that will be visually captured by the user. For the purpose of this research, a well implemented saliency toolbox was used [18] and the results were directly employed. While there is an evidence that saliency correlates well with spending human attention as a resource [19] (and hence workload), for the case of our study, we assume that the worker assesses a HIT by browsing (even mostly reading) the visible text in the HIT among other consideration. Therefore self intreduced test bases features (extracted from screenshot) were also considered. First, text was extracted from the image by binarizng the result of an OCR (Optical Character Recognition) operation provided in Matlab software package (release 7.1.0 were employed). Then, the image was binarized based on presence and absence of text in the image. Finally, the salient regions of the image and the text regions are merged in a binary image. The procedure is illustrated in Figure 2. Each image was divided into smaller windows and 13 features were extracted. In the following selected ones are described:

- **Text density:** A measure to express the size of the text in the window. It is the ratio of white pixels (text in the images) to the size of he window.
- **Threshold:** Average of the text density in all widnows.
- **Cumulative distribution of densities:** A set of values that represent how much of the image has text density below or equal to a certain level. This was done by just arranging the text density values in an ascending order. The levels considered were 25%, 50%,75%,95%, 98% and 100%.
- **Count of dense windows:** The number of *dense windows*. They are widnows that have more text in them then the threshold.
- **Largest connected region:** Connected regions were created using 8-connectivity neighborhood over teh dense windows. This quantity is stored in number of pixels as well as in number of windows.

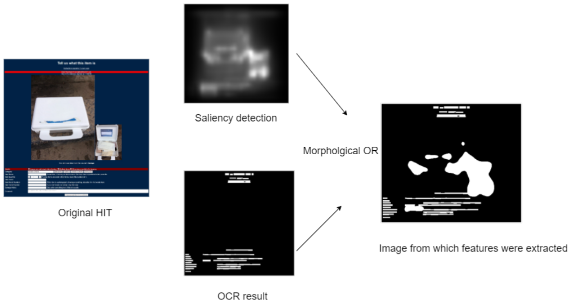


Figure 2. processing of HIT images for visual feature extraction LOW QUALITY

Table I. DISTRIBUTION OF DIFFERENT HIT TYPES IN THE FINAL DATASET.

Type	Count	Percentage
Content Access (CA)	13	3.8%
Content Creation (CC)	111	32.2%
Interpretation and Analysis (IA)	82	23.8%
Information Finding (IF)	68	19.7%
Surveys (SV)	37	10.7%
Verification and Validation (VV)	30	8.7%
Others	4	1.2%

III. RESULTS

A. *Data Screening*

Overall 6015 responses were collected containing 15 ratings per each HIT from the sample dataset. Answers from all workers were accepted. A two steps procedure was followed to remove unreliable responses before calculating the Overall Effort for each HIT. In the first step, each response was analyzed separately. As a result, 199 responses that have missing values and 439 responses that their workers specify they are unable to perform such a HIT were removed. In the second step, performance of each individual worker was evaluated. To do that, 30 % trimmed mean of Overall Effort was calculated for each HIT. The trimmed mean was chosen as the arithmetic mean is sensitive to extremely deviated ratings. Given that, average absolute deviation of ratings for each worker was calculated. Workers who had on average 25 or more deviation from the mean ratings were considered to be inaccurate. All responses from them (i.e. 597) were removed. Average deviation of 25 is chosen for cutting point as the number of suspected ratings start to exponentially raise for any cutting point below that. In addition 4 HITs were also removed from the dataset as there were less than 5 ratings per each remaining. Overall 387 HITs remained in the dataset with 4821 valid responses. Using all valid responses 30 % trimmed mean of Overall Effort was calculated for each HIT in the dataset.

Finally, the goodness of data cleaning procedure was evaluated based on the expected rank from the manipulated HITs in the dataset (c.f. Section II-A). From 14 manipulated HITs (each having 3 instance with different difficulty levels of *easy*, *medium*, and *hard*), 9 followed the expected order when ranking them based on the calculated Overall Effort value. For the rest, HITs with expected difficultly of *medium* were ranked nearly equal or with slightly more workload than their corresponding *hard* instances. Following the same procedure but using all responses (i.e. disregarding the data screening process) to calculate the Overall Effort, 2 out of 14 manipulated HITs follow the expected rank. As a result, the data cleaning procedure shows a satisfactory results.

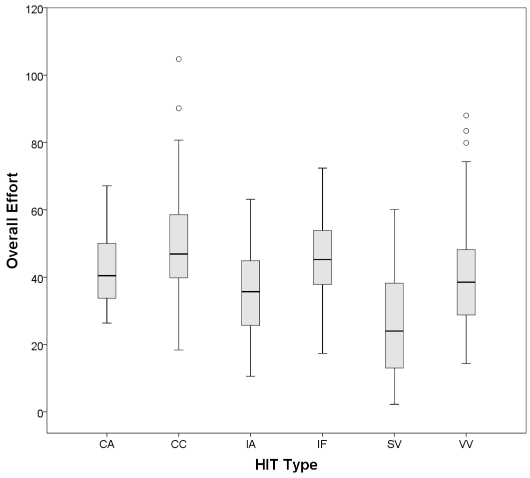


Figure 3. Distribution of *Overall Effort* of each HIT Type in the final dataset.

Removing the manipulated HITs, there remained 345 HITs in the dataset ready to be used for creating the prediction model. Distribution of HIT types in the final dataset are reported in Table I. Figure 3 illustrates the distribution of Overall Effort as well.

B. *Prediction of Overall Effort*

The remaining 345 HITs were then used to predict the trimmed mean overall effort for each HIT based on the extracted features. For this step, three different feature sets were used to train and test two prediction models one based on the Random Forest Regressor and the other based on the Linear Regression. A lot of the extracted syntactic features can be considered as noise, because of their low observation frequency in the HITs, therefore the model that used all available features exhibited a high standard deviation and can be considered as overfitted. To avoid overfitting, recursive feature elimination [20] in combination with the random forest regressor was used to determine the 20 most important features, which were then used as the second feature set we tested, which still had considerably high standard deviation. To obtain a third model, another 12 features were eliminated by discussing each features importance individually and testing discussed hypothesis manually. In an arbitrary order, the remaining and best performing features were: HIT Type, Reward, Number of Words, Number of Subclauses, Number of HTML tags: $<img>$, $<textarea>$, $<Input>$ (type: radio), and the 100% cumulative distribution of densities.
With a mean absolute error of 7.3, the resulting random forest model is a good estimator for overall effort. The relatively high standard deviation of 12.93 seems to be well within the expected deviation, because 258 out of 345 predictions were inside the 95 % confidence interval. Examining the importance of features assigned through random forest training shows, that reward is contributing the most information when predicting overall effort. The semantic features number of words and

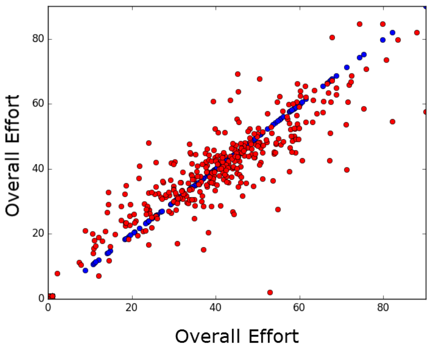


Figure 4. Overall effort random forest prediction results. Target values in blue, predicted in red.

Table II. Results of Overall Effort prediction with different models.

Prediction Model	r	MAE (std)	Overlapping 95%CI
Base Model	0	12.07 (10.15)	185
Random Forest final	0.75	7.30 (12.93)	258
Linear Regression	0.78	7.69 (6.23)	246

number of subclauses, as well as HIT type have similar significance and hold relevant information. The remaining features rank fairly low, which can be expected, because the number of syntactic appearances strongly depends on the design of the HIT and can therefore vary a lot.

The same process was repeated using Automatic Linear Modeling in IBM SPSS (version 24) to build a linear regression model. Full set of features were used in the forward stepwise [21] method to automatically select predictors. Marginal contribution of predictors, entered or removed from model, were evaluated using Akaikes Information Criterion Corrected (AICC) [22]. In addition, stability and accuracy of model was enhanced using Bagging (Bootstrap aggregating [23]) technique. That leads to an ensemble model with adjusted-$R^2 = 0.61$. Following predictors were used in the final model ordered by their importance: HIT Type, Subclauses, Rewards, 100%, 98%, 75%, and 95% cumulative distribution of densities, Number of Buttons, Number of unique stems, and Number of words. The ensemble linear regression model predicts the overall effort with mean absolute error of 7.69 ± 6.23. This model is also a good estimator of overall effort as 71% of predictions were inside the 95 % confidence interval.

Results for both provided models are reported in Table II including a based model which returns a constant value (the mean of overall effort calculated using all data in the dataset) despite the input.

C. Prediction of Acceptance

For the prediction of Acceptance of a HIT the responses from each individual worker was analyzed namely 4821 observations. The Acceptance value was first converted to a binary variable. From the dataset, the result of whether users would take a HIT or not was analyzed. This question was scored on a scale of 0 to 20. Any value on this score that is below 7 was labeled as a "not accepted" observation while the values above 13 were labeled as "accepted". For the rest of the values, the observations were simply removed. This created the new Acceptance variable. Three different models were assessed to predict the acceptance variable. The dependent variables were from meta-data i.e. Rewards, HITs available, and subjective assessments of workers i.e. Overall Effort, Complexity, Difficulty, Interestingness, Frequency (how often worker performs these kind of HITs), and ratio of reward to Overall Effort. For subjective variables both worker's specific ratings and mean values over all workers for that specific HIT were considered. In addition, observations that refer to the manipulated HITs were removed because the manipulated HITs did not have realistic value of reward parameter. The models were built using Binomial Logistic Regression, Random Forest, and Support Vector Machines (SVM) employing the training dataset that include 75% of observations (i.e. 2525). Later, they were evaluated with the test dataset (i.e. remaining 842 observations). The test results are reported in Table **??** The logistic regression model delivered an accuracy of 88.36% and was found to be significant with $\chi^2(4) = 1985.531, p < 0.005$ and Nagelkerke $R^2 = 0.727$. Besides the predictors mentioned in Table III, rest of the dependent variables were found be insignificant and were not used.

Table III. Logistic regression model

Predictors	co-effecients	std. error	significance ($p - value$)
Intercept	-1.863091	0.228881	3.95e-160
Reward to workload ratio	11.039114	3.994573	0.00572
Interesting	0.265450	0.013670	2e-16
Frequency	0.223550	0.012462	2e-16
Overall Effort	-0.045283	0.004906	2e-16

The accuracy of random forest model was 90.02% on the testing data. The model solved a classification problem and had 500 trees while the number of variables used for in each tree was 2. The Out-of-bag error rate of the model was found to be 10.57% for the training data set and the error was quite balanced over the two classes. Error for the not acceptance class was 10.7% while it was 10.5% for the acceptance class. Feature importance of the model was also analyzed using the GINI metric. *Frequency* and *Interestingness* were found to be most important variables while the *Overall Effort* and the *Reward to Workload Ratio* were relatively less important.

The accuracy of the SVM model using a linear kernel was 77 % with a precision of 95 %. The overall result did lack in significance, because with a low recall of only 56 %, many cases, where a worker would have accepted the HIT, were wrongly predicted as the worker not accepting it.

IV. Discussion

Even though the model for predicting the overall effort is performing reasonably well, better results could be achieved if the training data is more accurate. The study design highly affects the quality of the model and should therefore be designed in a way that allows for easy identification of participants that answer randomly.

Our approach in identifying such participants, as explained in the data screening section, was to remove participants from the dataset, if their answer deviated on average from the trimmed overall effort mean by 25, because on a scale from 0 to 150, which the overall effort was measured on, it accounts for 16.66% of the total scale. Therefore a standard deviation higher than 25 implies, that the HIT either can not be judged objectively or that participants submitted random answers.
The result of the feature selection stage also leaves room for further research, while Reward and syntactic features like number of input fields in examined HIT seem to be very important, semantic features i.e number of words and subclauses contribute significantly as well. This implies, that more sophisticated natural language processing might be applicable in extracting features that can be used to improve the accuracy of predicting the overall effort. The fact that reward is ranked the most important feature seems intuitive, because reward usually correlates with required effort. The importance of HIT type as a feature further implies, that different models could be used per HIT category to predict the overall effort.
For the prediction of acceptance the predictors behaved quite intuitively. Frequency correlated highly with the acceptance, which means that a worker that usually on HITs similar to the one in question, is more likely to accept a HIT. Similarly, If an HIT is considered interesting, it is more likely to be accepted than an un-interesting one. In the model overall effort is negatively correlated with the acceptance and therefore workers are less likely to take a HIT that requires more overall effort to be completed. These three predictors were found to be more significant than the predictor reward and reward to workload ratio. Reward to workload ratio, though a significant predictor, did not seem to be a better predictor in comparison to interesting” or ”overall effort”. This maybe due to the fact that HITs on MTurk platform are not priced as per a standard. Most of the HITs in our dataset are rewarded at $0.05 even though their workload effort and other aspects vary. A better result would be that a HIT with high reward to workload ratio would always be accepted.

V. Conclusion

VI. Some Informations for Writing

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as 3.5-inch disk drive.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: ”‘Wb/m2”’ or ”‘webers per square meter”’, not ”‘webers/m2”’. Spell units when they appear in text: ”‘...a few henries”’, not ”‘...a few H”’.
- Use a zero before decimal points: ”‘0.25”’, not ”‘.25”’. Use ”‘cm3”’, not cc. (bullet list)

C. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled. Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

$$a + b = c \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use ”‘(1)”’, not ”“”’Eq. (1)”’ or ”‘equation (1)”’, except at the beginning of a sentence: ”‘Equation (1) is ...”’

D. Some Common Mistakes

- The word data is plural, not singular.
- The subscript for the permeability of vacuum, and other common scientific constants, is zero with subscript formatting, not a lowercase letter ”‘o.”’
- In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an ”‘inset”’ not an ”‘insert”’. The word alternatively is preferred to the word ”‘alternately”’ (unless you really mean something that alternates).
- Do not use the word ”‘essentially”’ to mean ”‘approximately”’ or ”‘effectively”’.

- In your paper title, if the words ”‘that uses’” can accurately replace the word using, capitalize the ”‘u’”; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones ”‘affect’” and ”‘effect’”, ”‘complement’” and ”‘compliment’”, ”‘discreet’” and ”‘discrete’”, ”‘principal’” and ”‘principle’”.
- Do not confuse ”‘imply’” and ”‘infer’”.
- The prefix ”‘non’” is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the ”‘et’” in the Latin abbreviation ”‘et al.’”
- The abbreviation ”‘i.e.’” means ”‘that is,’” and the abbreviation ”‘e.g.’” means ”‘for example’”.

E. Figures and Tables

Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation ”‘Fig. 1’”, even at the beginning of a sentence.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity ”‘Magnetization’”, or ”‘Magnetization, M’”, not just ”‘M’”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write ”‘Magnetization (A/m)’” not just ”‘A/m’”. Do not label axes with a ratio of quantities and units. For example, write ”‘Temperature (K)’”, not ”‘Temperature/K’”.

VII. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank Mom and Dad.

References

[1] D. E. Difallah, M. Catasta, G. Demartini, P. G. Ipeirotis, and P. Cudré-Mauroux, “The dynamics of micro-task crowdsourcing: The case of amazon mturk,” in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW ’15. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2015, pp. 238–247. [Online]. Available: https://doi.org/10.1145/2736277.2741685

[2] J. Yang, J. Redi, G. DeMartini, and A. Bozzon, “Modeling task complexity in crowdsourcing,” in *Proceedings of The Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2016)*. AAAI, 2016, pp. 249–258.

[3] S. G. Hart, “Nasa-task load index (nasa-tlx); 20 years later,” in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage Publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[4] B. B. Bederson and A. J. Quinn, “Web workers unite! addressing challenges of online laborers,” in *CHI’11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2011, pp. 97–106.

[5] A. Felstiner, “Working the crowd: employment and labor law in the crowdsourcing industry,” *Berkeley Journal of Employment and Labor Law*, pp. 143–203, 2011.

[6] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, “The future of crowd work,” in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM Press, 2013, p. 1301. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2441776.2441923

[7] M. S. Silberman, “What’s fair? rational action and its residuals in an electronic market,” *Unpublished manuscript*, 2010. [Online]. Available: http://www.scribd.com/doc/86592724/Whats-Fair

[8] J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson, “Who are the crowdworkers?: shifting demographics in mechanical turk,” in *CHI’10 extended abstracts on Human factors in computing systems*. ACM, 2010, pp. 2863–2872.

[9] J. O’Neill and D. Martin, “Relationship-based business process crowdsourcing?” in *Human-Computer Interaction INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part IV*, P. Kotz, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds. Springer Berlin Heidelberg, 2013, pp. 429–446.

[10] W. W. Wierwille, M. Rahimi, and J. G. Casali, “Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 27, no. 5, pp. 489–502, 1985.

[11] S. M. Casner and B. F. Gore, “Measuring and evaluating workload: A primer,” 2010.

[12] D. De Waard, *The measurement of drivers’ mental workload*. Groningen University, Traffic Research Center Netherlands, 1996.

[13] G. B. Reid, C. A. Shingledecker, and F. T. Eggemeier, “Application of conjoint measurement to workload scale development,” in *Proceedings of the Human Factors Society Annual Meeting*, vol. 25, no. 1. Sage Publications Sage CA: Los Angeles, CA, 1981, pp. 522–526.

[14] S. G. Hart and L. E. Staveland, “Development of nasa-tlx (task load index): Results of empirical and theoretical research,” *Advances in psychology*, vol. 52, pp. 139–183, 1988.

[15] U. Gadiraju, R. Kawase, and S. Dietze, “A taxonomy of microtasks on the web,” in *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, ser. HT ’14. New York, NY, USA: ACM, 2014, pp. 218–223. [Online]. Available: http://doi.acm.org/10.1145/2631775.2631819

[16] B. Naderi, I. Wechsung, and S. Mller, “Effect of being observed on the reliability of responses in crowdsourcing micro-task platforms,” in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–2.

[17] S. Le Moan, A. Mansouri, J. Y. Hardeberg, and Y. Voisin, “Saliency for spectral image analysis,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 6, pp. 2472–2479, 2013.

[18] D. Walther and C. Koch, “Modeling attention to salient proto-objects,” *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[19] C. Shen and Q. Zhao, “Webpage saliency,” in *European Conference on Computer Vision*. Springer, 2014, pp. 33–46.

[20] J. Chen, “Enhanced recursive feature elimination,” *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*, 2007.

[21] M. Efroymson, “Multiple regression analysis,” *Mathematical methods for digital computers*, vol. 1, pp. 191–203, 1960.

[22] C. M. Hurvich and C.-L. Tsai, “Regression and time series model selection in small samples,” *Biometrika*, pp. 297–307, 1989.

[23] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.