Document Title	**Description of Statistical Tests and Procedures**
Project Title and acronym	DEtecting Stereotypes in human ComputAtioN Tasks (DESCANT)
Pillar	II. Sustainable RTDI System
Programme	Excellence Hubs
Grant Agreement	EXCELLENCE/0918/0086
Deliverable No.	D3.3
Work package No.	WP3
Work package title	Conceptual Framework
Authors (Name and Partner Institution)	E. Christoforou (CYENS)
Contributors (Name and Partner Institution)	J. Otterbacher (CYENS)
Reviewers	M. Kasinidou (OUC)
Status (D: draft; RD: revised draft; F: final)	F
File Name	D3.3_Statistical_Models_Bias.docx
Date	15 May 2020

This project has received funding from Research and Innovation Foundation (RIF) under grant agreement EXCELLENCE/0918/0086.

Draft Versions - History of Document				
Version	**Date**	**Authors / contributors**	**e-mail address**	**Notes / changes**
1.0	20/1/20	E. Christoforou	e.christoforou@cyens.org.cy	initiate document
1.1	5/3/20	E. Christoforou	e.christoforou@cyens.org.cy	update content
1.2	1/5/20	E. Christoforou	e.christoforou@cyens.org.cy	review version
1.3	15/5/20	J. Otterbacher	j.otterbacher@cyens.org.cy	final version

Abstract

This deliverable develops procedures for auditing the data produced from the crowdsourcing tasks by first laying down the possible types and format of the data collected. This initial step is followed by identifying possible working definitions of bias given the crowdsourced data and laying down the statistical test and methods to be used for identifying the existence of bias in the data.

Keyword(s):	Bias, Statistical Tests, Ground Truth, Crowdsourced Data

Contents

1. About this Deliverable

The final goal of WP3 is to develop procedures that will be valuable for us as requesters (i.e., users of crowdsourcing platforms, who submit a task) when pre-processing the crowdsourced data before being used by the machine learning algorithm (WP5). Additionally, in this deliverable, we define the basic concepts and propose some initial statistical models for identifying possible biases in our crowdsourced data.

2. General Micro-task Crowdsourcing Setting

Let's review once again the main crowdsourcing components:

We define **crowdsourcing** as an online participative activity in which a requesting entity (**requester**) proposes to a group of individuals (**crowdworkers** or **workers**) of varying knowledge and demographic background, via a dedicated **platform**, which connects requester with workers, the undertaking of a **task**. The worker agreeing and completing the task will receive a monetary reward and the requester will receive a response to the task. We refer to the set of processed task responses as **data**.

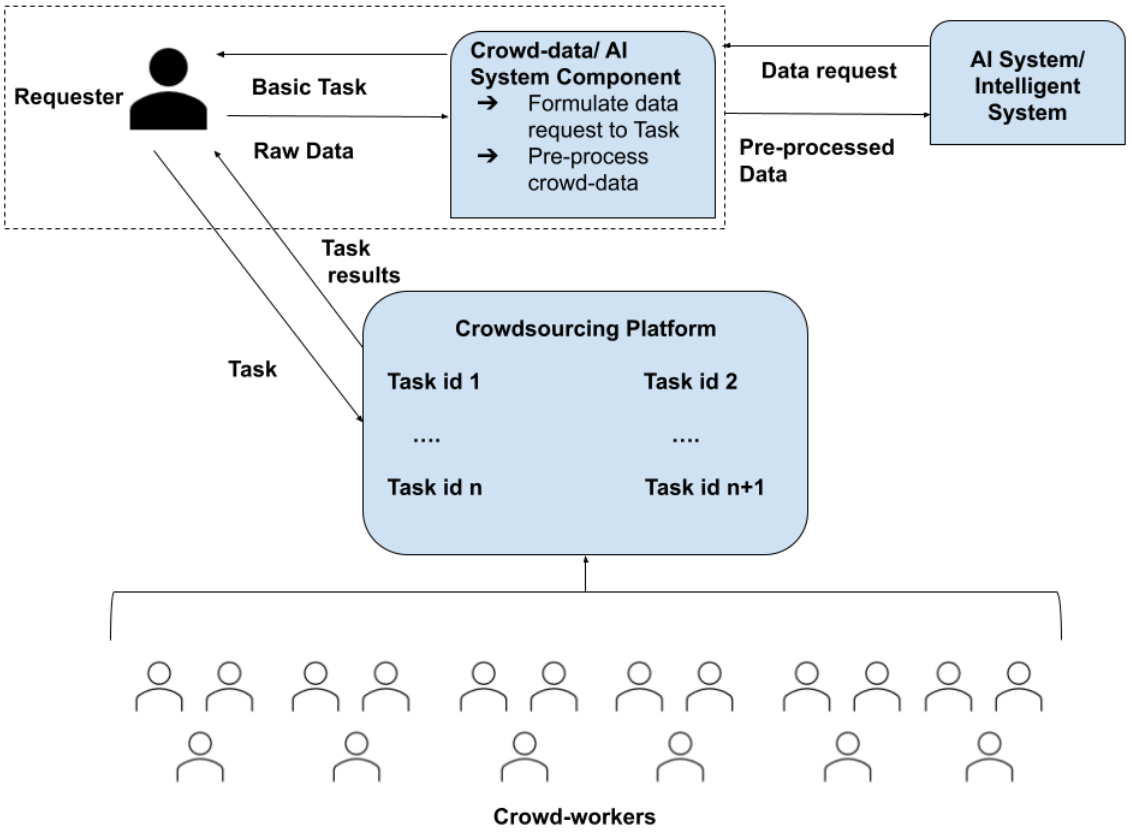

Figure 1. Crowdsourcing process and main actors when creating data for an AI system.

Recall, Figure 1., which presents an overview of the crowdsourcing process and its main actors when the purpose of crowdsourcing is the collection of data for an AI system (i.e., a system based on a machine learning process). In this deliverable, we will give particular emphasis to the pre-processing of the crowdsourced data task belonging to the Crowd-data/AI System component. Notice that the task of pre-processing the crowdsourced data depends on the data requirements and the raw data delivered to this component.

In the next section, we will review the possible types and formats of the crowdsourced data, defining in this way the components on which the statistical models for identifying bias will be based. Moreover, we will review some working definitions of bias that we can use during the analysis of the data and we will identify the possible limitations on the media artefact chosen when designing the crowdsourcing task. Finally, we will review based on a working example the possible statistical tests and methods used to identify bias in our crowdsourced data.

3. Auditing the crowdsourced data

3.1 Type and format of crowdsourced data

The crowdsourcing task will be based on a media artefact (i.e., video, audio, image etc.) which the crowdworkers will be asked to annotate following some instructions. Additionally, given our finding in the Conceptual Framework D3.2, we will either have some demographics on the crowdworkers available and/or we will ask crowdworkers to self-report demographics (i.e., age, gender, race etc.).

Each crowdworker responding to our task will generate a vector of responses for that task matching the vector of annotation questions and demographic questions asked. Each task will refer to one or multiple media artefacts of the same type (i.e., images on the same topic). The vectors of responses provided by a crowdworker on a task are coupled with the media artefacts shown to the worker. Each media artefact will be annotated by more than one crowdworker, following our findings in D3.2. Redundancy is an important tool when looking to improve the quality of the collected data or identify the effect of bias on the annotations of a media artefact (but also in general for many other types of tasks (Gadiraju, Kawase and Dietze 2014)).

To facilitate the discussion, we will assume the working model depicted in Figure 2., that is easily applicable to different types and ways of collecting crowdsourced data. Data collected for an image annotation task can be viewed in many ways. Naturally, when the collected data will be provided to a Machine Learning (ML) algorithm that needs to predict the features of the image, each image will be accompanied in the scenario of Figure 2 (left side), by three annotations per feature, as many as the crowdworkers who were asked to annotate a single image. Given the objectives during the ML process and the actual options for annotating each feature (i.e., multiple choice answers, open form answer with no predefined selection) the data will be aggregated.

As we explained in D3.1 and D3.2 this aggregation process of the data might allow the propagation of biases into the ML algorithm. On the right side of Figure 2. we present a possible view of the data that will help us identify and demonstrate the existence of biases in the collected dataset. As we can see from Figure 2., the data can be analysed based on the answers provided by

each demographic group (i.e., women or participants older than 50 years) and per each of the requested features.

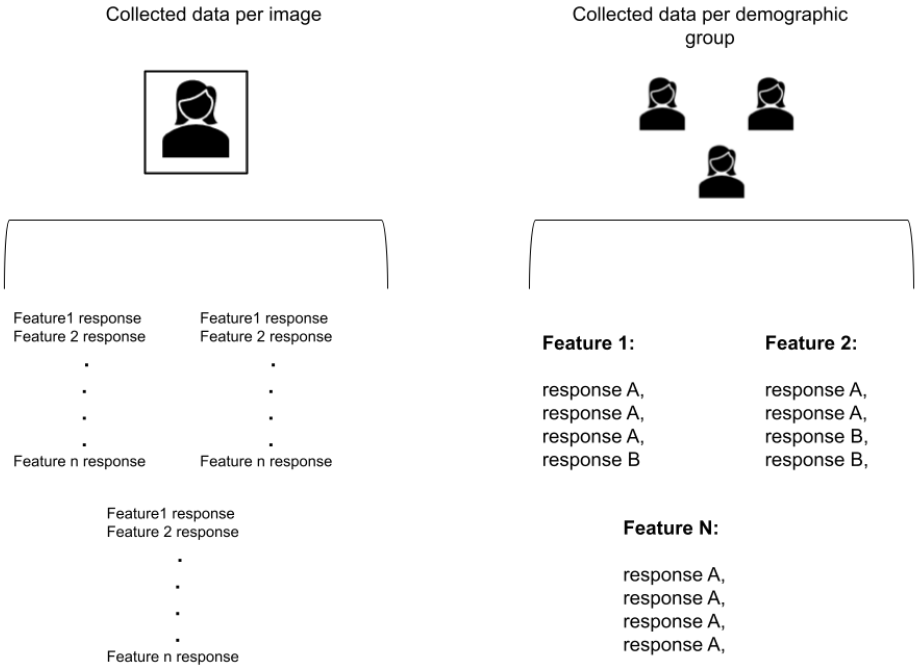

Figure 2. Different views of the crowdsourced data. Left, data view for the ML algorithm, Right, view of the data suitable for analysis of biases.

3.2 Working definition of bias

Through the analysis of our crowdsourced data, we would like to identify whether and to what extent those data include human biases. In the context of our study, human biases arise from crowdworkers providing subjective judgements on subjective tasks because they might suffer from cognitive biases (Eickhoff 2018), creating their own "subjective social reality" systematically deviating from ***norm in judgement***. Additionally, crowd workers may suffer from stereotypical beliefs about the characteristics, the attributes and the behaviors present in a certain group of people (Hilton and Von Hippel 1996). When a requester taps into a large pool of workers from around the globe, it is hard to identify the workers' stereotypical beliefs that can span from gender to ethnicity, sexual orientation, and religion to less apparent stereotypes that are regional or that are based on a person's socio-economic status or that are related to a person's profession or educational background.

Considering the working example of Section 3.1 above, we consider the following working definitions of bias:

We identify as bias the ***systematic deviation*** of the responses of a group of crowdworkers, sharing a common demographic characteristic, ***from the norm*** (the ground truth).

An example of bias is the following finding, in a given dataset of crowd responses: all male crowdworkers annotating images of male and female subjects with the perceived age of the depicted person, systematically providing the correct age of male subjects with a deviation of ± 3 years, while reporting the correct age of female subjects with a deviation of ±7 years.

We hypothesize the existence of bias among the responses of two different groups of crowdworkers (each sharing a common, comparable demographic characteristic) when the distribution on the responses provided by the two groups varies significantly.

For example, when the subset of Asian crowdworkers, as compared to the subset of Caucasian crowdworkers, annotates significantly differently the perceived race of the subjects in the depicted images, we can hypothesize the existence of bias.

We hypothesize on the existence of bias when a media artefact (i.e., image) grouped according to a certain demographic characteristic (i.e., race) are not annotated in the same way (i.e., same frequency of features).

For instance, when examining the whole set of responses from the crowdworkers, we identify that images of Black subjects received more annotations referring to the perceived health condition of the subject, as compared to images of White subjects.

3.3 Benchmark for the produced data and ground truth

As we have seen from the working definition of bias, in order to be in a position to identify or hypothesize the existence of bias, we either have to have a *ground truth* with which we compare the received crowdworker responses, or a way to *benchmark* what is expected or acceptable against the crowdworkers' responses.

Seizing a set of media artefacts on which the annotation task will be based, is a process that must be carefully planned. Many media artefacts datasets exist, for example on Kaggle[1], that are of particular interest to this study, especially for the potential expression of social stereotypes that we aim to study. However, one of the main considerations when choosing a dataset as our media artefact, must be the existence or not of a ground truth on at least some of the features of the dataset. For example, an image dataset where the race of the depicted subjects is included as part of the dataset and arrives from the depicted subjects self-reporting their race, is a more desirable dataset as compared to a dataset where the race feature is the result of expert annotators agreeing on the race of the depicted subject.

3.4 Statistical models capturing bias

Considering the running example we have used thus far, let's assume that in each crowdsourcing task we consider a single image and for each image we ask three distinct workers to provide us

[1] https://www.kaggle.com/

with ten, open-ended single-word tags that describe that image. Thus, for each image we will have in total, 30 tags. In this case, we could pre-process our data, identifying the different categories that the reported tags belong to. If we wanted to observe, for example, if the gender of the crowdworker correlates to the way the worker chooses the reported tags, we could apply a chi-square test to examine whether the number of occurrences for the category of tags we want to study depends on the gender of the crowdworker. Looking at a specific category of tags, we can use the two proportion z-test to see whether the proportion of tags used that are of that particular category are the same across genders.

Another likely scenario for analysis is if we wanted to test the relationship between a factor (e.g., the socio-economic status of the crowdworker) and the accuracy of the annotation features provided per task, by the workers. Here, we could use the one-way ANOVA test, testing first the homogeneity of variances using the Barlett's test. If the test yields a statistically significant result (i.e., with a p-value < 0.05, indicating that the factor of interest and the observed accuracy are not independent of one another), we move forward with the ANOVA test. After observing a statistically significant result here, we can test for the normality of the residuals using the Shapiro test.

References

Chang, Joseph Chee, Saleema Amershi, and Ece Kamar. 2017. "Revolt: Collaborative crowdsourcing for labeling machine learning datasets." *In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* 2334-2346.

Eickhoff, Carsten. 2018. "Cognitive biases in crowdsourcing." *In Proceedings of the eleventh ACM international conference on web search and data mining.* 162-170.

Gadiraju, Ujwal, Ricardo Kawase, and Stefan Dietze. 2014. "A taxonomy of microtasks on the web." *In Proceedings of the 25th ACM conference on Hypertext and social media.* 218-223.

Hilton, James L., and William Von Hippel. 1996. "Stereotypes." *Annual review of psychology* 47 (1): 237-271.

McGarty, Craig Ed, Vincent Y. Yzerbyt, and Russell Ed Spears. 2002. "Stereotypes as explanations: The formation of meaningful beliefs about social groups." *Cambridge University Press.*