Socio-Economic Diversity in Human Annotations

Shaoyang Fan
The University of Queensland
Australia
fsysean@gmail.com

Pinar Barlas
CYENS Centre of Excellence
Cyprus
p.barlas@cyens.org.cy

Evgenia Christoforou
CYENS Centre of Excellence
Cyprus
e.christoforou@cyens.org.cy

Jahna Otterbacher
CYENS Centre of Excellence
Cyprus
j.otterbacher@cyens.org.cy

Shazia Sadiq
The University of Queensland
Australia
shazia@itee.uq.edu.au

Gianluca Demartini
The University of Queensland
Australia
g.demartini@uq.edu.au

ABSTRACT

Human annotations can help indexing digital resources as well as improving search and recommendation systems. Human annotators may carry their bias and stereotypes in the labels they create when annotating digital content. These are then reflected in machine learning models trained with such data. The result is a reinforcement loop where end-users are pushed stereotypical content by the search and recommendation systems they use on a daily basis. In order to break the loop, the impact on models of using diverse data that can better represent a diverse population has been looked at.

In this work, we look at how human annotators in the US annotate digital content different from content which is popular on the Web and social media. We present the results of a controlled user study in which participants are asked to annotate videos of common tasks performed by people from various socio-economic backgrounds around the world. We test for the presence of social stereotypes and investigate the diversity of the provided annotations, especially since some abstract labels may reveal information about annotators' emotional state and judgment. We observe different types of annotations for content from different socio-economic levels. Furthermore, we find regional and income level biases in annotation sentiment.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

crowdsourcing,stereotypes,video annotation

ACM Reference Format:
Shaoyang Fan, Pinar Barlas, Evgenia Christoforou, Jahna Otterbacher, Shazia Sadiq, and Gianluca Demartini. 2022. Socio-Economic Diversity in Human Annotations. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3501247.3531588

WebSci '22, June 26–29, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9191-7/22/06…$15.00
https://doi.org/10.1145/3501247.3531588

1 INTRODUCTION

With the growth of social media, a massive amount of digital content is created every day by people over the Web. Algorithms are utilised to facilitate the retrieval and recommendation of such digital content. However, these algorithms can alter how people perceive the world, and their social implications are being extensively debated in academia. For instance, search engine algorithms have frequently been accused of social bias. Kay et al. [22] showed that search algorithms reinforced gender stereotypes via image search queries about professions. Users may see men appearing in search results for doctors but women in search results for nurses. This may affect young girls' perceptions of their future career prospects and gender expectations [27]. A study of job recommendation systems showed how setting the gender to female resulted in fewer ads for higher-paying jobs [8].

Annotations, also known as tags or labels, are essential parts of search algorithms and recommendation systems. Annotations can, for example, be utilised to index multimedia data to help people retrieve the desired content through queries [36]. The features extracted from annotations can also be applied to rank search results and to improve search and recommendation performance [13]. Crowdsourcing can be leveraged to efficiently collect large amounts of annotations for digital content. Along with tagging objects and actions represented in the digital content, human annotations may reveal information about the annotator's emotional state. These are possibly their perceptions and beliefs regarding the digital content and the scenes and/or persons depicted in it. Related to these human annotation processes and focussing specifically on short videos, our Research Questions (RQs) are as follows:

- RQ1: What video characteristics are commonly highlighted by human annotators in the labels they provide?
- RQ2: Is there any correlation between human annotation behaviour, their demographics, the video content, and the type of annotations they use?

Annotations created by users can reflect how they describe and view digital resources [23], but annotators may unconsciously project social stereotypes in the annotations they provide. Barlas et al. [3] found that computer vision algorithms trained with human annotations struggled with gender recognition, particularly when dealing with images of women, people of colour, and non-binary individuals. They discovered that this occurs because the training data inadvertently associate gender with socially stereotyped

scenario occupations. While gender and race stereotypes in annotations have been studied [3, 21], there is still a lack of research on bias in annotations from the perspective of socioeconomic status (SES). If SES stereotypes in video annotations are not adequately studied, video search engines optimised with these annotations may reinforce these stereotypes [29], resulting in social inequality [10]. Filling this research gap is the second contribution of our work. Our RQs also include:

- RQ3: How are users' biases and/or stereotypical beliefs reflected when annotating SES-diverse content?
- RQ4: Can we observe human annotators' social class competence from post-annotation questions?

To answer these RQs, we perform controlled user studies utilizing crowdsourcing as a recruitment mechanism where we ask participants to label SES-diverse content to understand their perception of the presented video content. Our findings support the following conclusions. The most frequently used annotations were general and abstract (RQ1). The order in which annotations are entered influences their type and the number of videos from particular SESs correlates with the number of distinct annotation types (RQ2). We observe regional and income-level biases when performing sentiment analysis on abstract annotations and analysing the post-annotation questions (RQ3 and RQ4).

2 RELATED WORK

Multimedia annotations. Søbak and Pharo [35] analysed the annotations created by professional production staff on television programs. They discovered that television production experts with extensive domain knowledge do not compensate for a lack of indexing training and motivation. On the other hand, cost and scalability constraints limit the amount of expert annotation that can be collected, and experts may also have a different understanding of video content than the general public [12]. Automatic annotation solves the problem of scalability but is restrained by the types of available annotations. Currently, automatically generated annotations are mainly about objects in a video, and the semantic gap problem still exists [36]. Additionally, automated annotation systems exhibit unexpected biases exacerbated by a lack of diverse training data [30, 32].

Social annotations leverage a diverse set of users by allowing them to freely annotate videos with keywords relevant to their interests, resulting in significantly increased annotation efficiency. The quality of crowd-generated annotations cannot be guaranteed, and these annotations may be influenced by diverse factors [1]. For example, some YouTube users employ many "Click-Bait" annotations to promote their digital media. Obtaining annotations via crowdsourcing platforms such as Amazon Mechanical Turk (MTurk)[1] can address the incentive issues mentioned above. There is no direct relationship between videos and crowd annotators, so exaggerating annotations to boost search rank for videos does not happen. Moreover, quality control methods can be employed to control the quality of their output annotations. In this work, we focus our study on annotations generated by crowd workers.

[1] http://www.mturk.com/

There are few studies on the diversity of video annotations generated by crowd workers. Some studies have explored how to use video labeling games to annotate movie clips [12] and TV shows [15, 19], but the number of annotations they generate and the number of video clips they use are small. More comparisons between our study and these past studies are presented in Section 4.

Economic Inequality. Research in social psychology has looked at how inequality can negatively impact people. For example, Kraus et al. [24] explain that economic inequality affects people daily through the communication of class signals. In social media, this could be someone's appearance (e.g., clothing, tangible items), communication style, or aspects of their cultural preferences and choices (e.g., leisure activities). The potentially harmful consequences from the continual signalling of the social class include: i) sorting people out into groups; ii) perpetuating stereotypes; iii) class conflict. They also describe class signals as being the "tinder for class conflict", as they impact people psychologically, in particular when people are in a position - as they are on social media - to make social comparisons. In our work, we conduct a controlled user study in which participants annotate content across income levels and regions and determine whether unconsciously projected social stereotypes were evident in the annotations they provided. This is a first step towards developing new user-centric web search and recommendation algorithms capable of leveraging users' emotional states and well-being via abstract annotations, maximising positive emotions and minimising harmful social comparisons.

3 STUDY SETUP

3.1 Study Design

In order to collect and analyze the annotation behaviour of crowd workers, we leverage MTurk's API to build an online crowdsourcing annotation platform. We recruit participants and send rewards through the MTurk platform while the task display and worker responses are independent of the MTurk platform. A separate database on a dedicated server stores the workers' data (i.e., behaviour logs and annotations).

Task page. MTurk workers were first required to complete a questionnaire to answer background questions, such as gender, age, family income and education level. Next, they were required to watch four short videos in sequence and provide eight annotations for each video. Our pilot tests found eight annotations to be the optimal number of annotations to collect for the short video we used, in terms of data quality. To analyze annotation diversity, the type of annotations to be entered by workers was not specified in the task instructions. Figure 1 displays the task interface for the annotation phase with the video on the left and the annotation section on the right. After completing the annotation phase, for all four videos, workers were asked to categorize each video through four follow-up questions: 1) In what region of the world do you believe the video was taken? (options include Africa, Asia, Europe, the Americas); 2) Relative to this particular region, which socioeconomic class would you say this person/household belongs to? (options include Lower class, Lower-Middle class, Upper-Middle class, and Upper class); 3) Would this video be appropriate as a result for a Google video search on 'hand washing'? 4) Would this

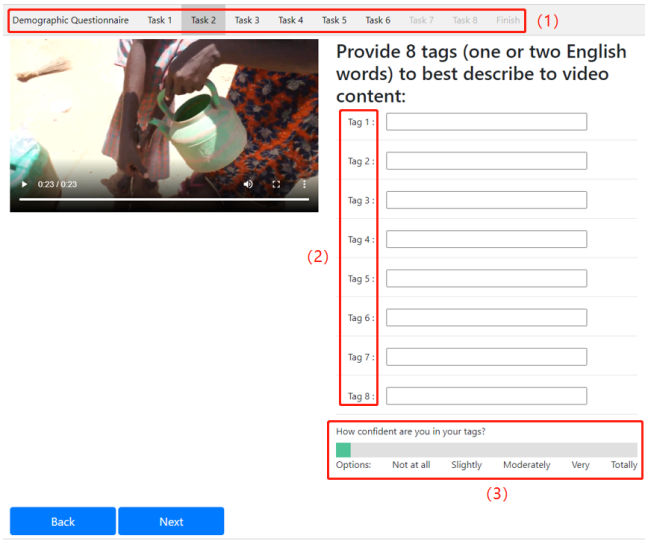

Figure 1: Screenshot of the annotation task: (1) tabs and buttons allow workers to revise answers; (2) workers enter annotations sequentially; (3) a 5-point Likert scale is used to report the self-perceived confidence in answering questions.

video be suitable for a COVID-19 hand hygiene service announcement? (options for the last two questions were on the 5-point Likert scale from 'Not at all' to 'Totally'). The screenshot for follow up questions and aligned options are shown in 2. On completing all tasks, workers were rewarded via MTurk.

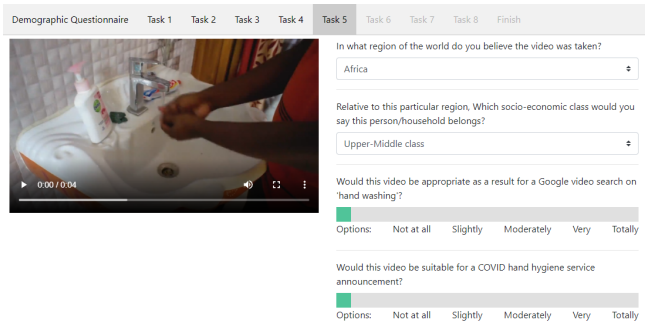

Figure 2: The screenshot of follow up questions page.

Video Selection. All short videos we employed in our study are published under a Creative Commons license 4.0 by Dollar Street [14]. Dollar Street is a website that provides a way to explore the daily lives of hundreds of families of different income levels worldwide through videos and pictures, aiming to overcome the media's skewed selection of digital content. For each family, the photographer takes pictures of multiple household items such as a toothbrush or a pair of shoes and records videos related to everyday life activities, such as brushing teeth, eating, and doing laundry. For our study, we selected only one topic depicting one daily activity, *hand washing*, which is performed in different parts of the world, in households with diverse income levels.

The experimental factors we used to select the videos used in our study were region and income level. We chose videos from four geographical regions: Africa, Asia, Europe and the Americas and we also selected videos based on four different income levels for each region. Income on Dollar Street is measured in US dollars adjusted for purchasing power parity rather than salary or income. For example, if a household consumes the corn they grow every month, that corn's value is factored into the household's total income. The income level for each region is calculated based on the quartiles of income within that region. We selected 7 videos for each of the 16 condition (4 regions × 4 income levels), 112 videos in total. We ensure that the number of videos is balanced across situations. When workers watch the videos, income and region in the video are not displayed. We excluded videos from the United States specifically as the annotators were recruited from the United States and we did not want to assume their familiarity with some of the content. The mean video duration is 13.7 seconds ($SD = 9.14$ seconds).

Quality control. To observe different annotation behaviours and because there is no ground truth in the data, our study does not directly use gold questions to control quality also because they are vulnerable to be attacked [6]. However, using our logger system, we employed four different quality control mechanisms to remove low-quality annotations from the final dataset. We tracked all workers' activities, such as mouse-clicking and text inputting with the corresponding timestamp. Firstly, we determined whether workers played the video or not before annotating it; cases where annotations were provided without watching the video were automatically classified as annotations to be discarded. We also use task time to control for quality. The minimum task duration time for each video should be greater than the length of the video, where task time is calculated by whether workers stay on the task page, preventing workers from prolonging the time by performing other tasks. Furthermore, if a worker produces mostly illogical or random text (over 50% for each video), their data is removed. In addition, workers' IP addresses were recorded in order to prevent them from using multiple MTurk accounts. Any worker who fails to pass one of the quality control checks is not exposed to future tasks. Tasks with low-quality annotations are republished to ensure that each video receives the same number of high-quality annotations.

Participants. Each video is annotated by 10 workers who have passed quality checks, and each worker annotated videos from four different families (present in Dollar Street) in the same region but with varying income levels. Workers are shown these four, non-repeating videos in random order. A worker can only perform one task on our platform. As a result, we recruited 280 qualified workers through MTurk. As the United States is the primary source of workers on MTurk [33], this research only recruited US-based workers. Restricting ourselves to US-based subjects may be a limitation of the study, but it also gives us a more controlled setting (drawing subjects from the same population) for better understanding interaction behaviour with SES-diverse content. Participants were paid $1.50 for completing four short video annotations. The rate of $1.50 is based on the platform's market price and takes the average task time in a pilot experiment into account (about 10 minutes).

The University of Queensland' ethics committee reviewed and approved the task, experimental design, and data collection before conducting this study. Prior to starting the experiment, workers gave informed consent to how their data would be collected and used in the study.

3.2 Classification of Annotations

We classify video annotations into three top-level categories based on Hollink et al. [18]'s annotation classification system: non-visual, perceptual, and conceptual. The non-visual level annotations describe the video's context, such as title and video length, rather than its content. The perceptual annotations are derived from the video's audio features, such as volume level, and visual features, such as colour. The Panofsky-Shatford mode/facet matrix [2] is utilised to represent the semantic content of the video in conceptual level annotations. The conceptual annotations are divided into three levels in this matrix model: general, specific, and abstract [31], and each level is further subdivided into four aspects: who, what, where, and when [34]. Because we collected a small number of annotations that could not be classified as described above, we created an 'Other' category to classify these annotations. A worker is excluded from the dataset if most of their output (over 50% of each video) consists of nonsensical words. Table 1 displays the definition of annotation categorization and annotation examples across the Hollink's and Panofsky-Shatford models.

Preprocessing. We pre-processed the 8,960 annotations we collected before classifying them. First, obvious spelling errors were corrected. For example, 'buket' can be easily corrected to 'bucket' depending on the context of the video. Then, we assigned base forms to those annotations using an English Lemmatizer pipeline (i.e., the *en_core_web_trf model* (roberta-base) by the Huggingface library [2]. As a result, 3,120 unique annotations were generated.

Manual coding. We manually classified annotations as, compared to automated categorization approaches, a small group of researchers manually annotating the dataset can minimize bias in the process [4]. When manually classifying annotations, we adhere to the following guidelines. Firstly, an annotation can only belong to one category. We encode annotations with the class that best represents them. An object or action with a subjective description is abstract, such as "beautiful lady", whereas "lady" is a general annotation. A person name, "Mary", falls under the a Specific annotation category. This rule corresponds to Gligorov et al.'s classification method [15]. For the four aspects: the Who represents the video clip's subject (person, object, etc.); the What represents an action or event in the video; the Where refers to a location; and the When represents the time. Furthermore, an annotation, such as "white", may be ambiguous. In that case, we examine the video to determine whether "white" is used to describe the video's background, the people, or an object in the video. Another difficult annotation example is the tag "no soap" which is classified as "abstract/What" whereas "soap" is classified as "general/Who". The addition of "no" to the annotation makes a key difference as the soap does not appear as an object in the video but rather it is a judgment from the annotator. If some annotations result in disagreement among researchers after being observed through the video, they were classified as Other. Two researchers (coders) have coded individually 400 randomly selected annotations, resulting in high agreement (Cohen's $k = 0.97$). The rest of the annotations were coded in the same manner by one of the coders.

[2] https://huggingface.co/

4 RESULTS

4.1 Annotator Background and Behavior

A total of 523 annotators were recruited through MTurk to participate in our experiment. Table 2 depicts the number and proportion of annotators who completed the task, abandoned the task, and failed the quality check. 210 annotators chose to abandon the task before its completion, and 33 people failed the quality test. The percentage of people (40.15%) who gave up on the task after starting is consistent with previous research [17]. Analysing the behaviour of annotators who abandoned the task, we discovered that 146 annotators chose to abandon on the first video page, accounting for 70% of total abandonment, while 34 annotators (16.2%) chose to abandon on the second video page. Annotators may have chosen to give up because the monetary reward was not appealing enough to them given the task, or the video content caused them to give up because they were uninterested.

We examined video attributes on abandoned task pages to see if two hidden variables of video content, region and income level, influenced annotators' decision to abandon the task at that point. We observed that the number of abandoned videos was distributed evenly across regions and income levels (mean number of videos abandoned at different income levels was 52.5 with a SD of 2.29; mean number of videos abandoned across regions was 52.5 with a SD of 9.55). A two-way ANOVA was conducted to examine the effects of region and income level on the number of abandoned videos, and there was no statistically significant difference for different income levels and regions, with all $p > 0.05$.

Males made up 55% of the participants in our study, and median age was 36 years. Participants are well-educated because over 64.29% of annotators possess a four-year college degree (66.88% for male and 61.11% for female). This rate is consistent with previous studies [26]. 64.99% of annotators earn less than $60,000 per year (70.13% for male and 58.73% for female).

4.2 Characteristics of Annotations

In order to sufficiently understand the types of annotations that users typically employ, we conduct a qualitative study of the obtained user annotations to understand the relationship between the content described in the videos and the types of annotations used for these descriptions. The number and percentage of tags across the categories of the Hollink model are shown in Table 3. Among the 8,960 annotations, only 0.38% of annotations are about the non-visual level, and these annotations are mainly about the description of the video's context, such as video length and type. Around 1.55% of the annotations are related to the perceptual level, describing the video's colour and sound. The majority of the annotations (95.59%) are conceptual descriptions of objects and actions observed in the video, while the remaining 2.48% are unclassifiable.

According to the Panofsky-Shatford model, the Conceptual-level annotations are classified as per Table 4. Most annotations belong

Hollink's model	Panofsky-Shatford matrix	Who	What	Where	When
Conceptual	Specific	Individually named person, object (Covid-19, Indian, Crocs)	Individually named events (2018 world cup*)	Individually named location (Africa, Asia, Haiti)	Specific time (hand washing day)
	Abstract	Object with subjectivity (clean hand, unclean water)	Emotions, relationship, judgment (poor, dirty, no soap)	Symbolised place (clear area, poor place)	Symbolsed time (family time*)
	General	Person, object (soap, woman, sink)	Action, event (hand wash, rub, use soap)	Location (rural, bath room)	Cyclical time (morning)
Nonvisual	The annotations are intended to describe the context of the video rather than the content. (20 seconds, related playlist, short video)				
Perceptual	The annotations are about descriptions of visual features like colour or descriptions of audio features like volume. (yellow, loud, relaxed noise)				
Other	These annotations do not fall into all of the above categories. They may include random input, pure prepositions, irrelevant words. (se, in, none)				

Table 1: The definition of annotation categorization with examples in terms of Hollink's model and the Panofsky-Shatford model. * represents no example of annotation at this category in our study.

Completion	Abandonment	Failure
280 (53.54%)	210 (40.15%)	33 (6.3%))

Table 2: Annotator behavior rates (the number of annotators and percentage).

Hollink Model	The Number of Tags
Perceptual Level	139 (1.55%)
Nonvisual Level	34 (0.38%)
Conceptual Level	8565 (95.59%)
Other	222 (2.48%)
TOTAL	8960 (100.0%)

Table 3: Distribution of the tags across the categories of the Hollink model

	Specific	Abstract	General	TOTAL
Who	31 (0.36%)	827 (9.66%)	3163 (36.93%)	4021 (46.95%)
What	0 (0.0%)	1890 (22.07%)	2265 (26.44%)	4155 (48.51%)
Where	15 (0.18%)	64 (0.75%)	198 (2.31%)	277 (3.23%)
When	1 (0.01%)	10 (0.12%)	101 (1.18%)	112 (1.31%)
TOTAL	47 (0.55%)	2791 (32.59%)	5727 (66.87%)	8565 (100.0%)

Table 4: Distribution of the conceptual-level annotations across the categories of the Panofsky-Shatford model.

to the What facet (48.51%) and the Who facet (46.95%), while the Where and When facets contain significantly fewer annotations. For the total number of annotations at various abstraction levels, most annotations (66.87%) are general, while 32.59% of annotations are abstract, and only 0.55% are at the specific level. The relationships between abstraction levels and facets indicate that the vast majority of annotations in the Who facet are general (e.g., "Bar soap"),

	Specific	Abstract	General	TOTAL
Who	17 (0.7%)	250 (10.31%)	645 (26.59%)	912 (37.59%)
What	0 (0.0%)	760 (31.33%)	613 (25.27%)	1373 (56.6%)
Where	3 (0.12%)	45 (1.85%)	62 (2.56%)	110 (4.53%)
When	1 (0.04%)	4 (0.16%)	26 (1.07%)	31 (1.28%)
TOTAL	21 (0.87%)	1059 (43.65%)	1346 (55.48%)	2426 (100.0%)

Table 5: Distribution of non-repeated annotations across the categories of the Panofsky-Shatford model.

occasionally abstract (e.g., "pandemic"), and infrequently specific (e.g., "COVID 19"). However, in the What facet, the descriptions are both general (e.g., "hand washing") and abstract (e.g., "good hand wash") but never specific. In the Where and When facet, most of the annotations are generic (e.g., "bathroom", "after toilet").

The annotation occurrence frequency is an essential indicator in some annotation systems to control quality, with non-repeated terms being removed as low-quality annotations [15]. Infrequently used annotations, on the other hand, can be instrumental to understand annotation behaviour [23]. For example, Chan observed that less commonly used annotations can significantly enrich the context of an image and increase accessibility [5]. Furthermore, Eleta and Golbeck found that less widely used annotations were more likely to reflect specific cultural contexts and increase the number of visitors to the annotated content [11]. Among the annotations we collected, 2,644 annotations that appeared only once accounted for 29.5% of the total annotations. We discovered that, compared to the distribution of all annotations, conceptual level annotations still account for the significant part with 2,426 (91.75%), while the percentage of annotations that could not be defined increased to 5.37% with 142 annotations. These non-repeated annotations at the conceptual level are classified according to the Panofsky-Shatford model presented in Table 5. Comparing the total number of non-repeated annotations at various abstraction levels with all annotations, the ratio of general annotations decreases from 66.87% to 55.48%, while the proportion of abstract annotations increases from

32.59% to 43.65%. When the total number of infrequently used annotations in each facet is compared to all annotations, the ratio of annotations in the Who facet decreased from 46.95% to 37.59%, but the proportion of annotations contained in the What facet increased from 48.51% to 56.6%. In addition, on the What facet, the proportion of less common annotations at abstract level to all annotations increased from 22.07% to 31.33%.

Discussion. Our findings are consistent with previous studies using the same classification method [12, 15, 19, 23]. To begin, the majority of crowdsourced descriptions are at the conceptual level. Second, most conceptual annotations belong to the general type. The difference between this study and previous ones is the proportion of abstract annotations to the total number of conceptual annotations. Gligorov et al.'s analysis of the annotations associated with reality television shows (long videos) revealed that the number of specific annotations was higher than the number of abstract annotations [15]. A possible explanation for this difference could be the presence of subtitles in the videos they used in the experiment, which provide precise information to human annotators.

However, our findings are consistent with Estrada et al.'s study, and the medium they used is also short videos (movie clips). We both observed that general/Who (e.g., "plate") and general/What (e.g., "use soaps") dominate the set of annotations. This indicates that in the annotation of short videos, people are more inclined to give broad descriptions of the subjects in the scenes and the actions that occur. These types of annotations can assist users in retrieving the video and provide context for the video prior to watching it. Additionally, these annotation categories can be used as pre-training datasets in machine learning to enhance models' ability to detect objects and actions in the video.

We also have consistent finding with Estrada et al. [12], in that the third most frequently used annotation is abstract/What (e.g., "neatness"), as people use abstract terms to describe events or actions in a scene, such as emotions and judgment. Abstract annotations are subjective in nature and are used to ascertain what other users think about a video [23]. Abstract annotations enable users to discover videos that share some common interests. Many abstract types of annotations, for example, show the potential for sentiment search. People can improve their viewing experience by finding videos that resonate with them emotionally more easily in the recommendation system. Additionally, abstract terms can assist users in locating words that accurately describe their inner feelings. For example, the search system will provide the user with these abstract terms to convey their emotional needs adequately.

Motivating users to provide abstract annotations is critical because of their importance. According to Estrada et al., increasing the variety of annotations is easier when annotators are given example annotations on guidelines [12]. On the other hand, Gligorov et al.'s study omitted many potentially useful annotations, such as abstract annotations, by focusing solely on high-frequency annotations [15]. We find that the proportion of abstract annotations is higher for low-frequency annotations. Indeed, these infrequently used annotations are more likely to remain stable over the annotation set's lifetime than high-frequency annotations [16], allowing for the presence of minority viewpoints. Additionally, less frequently used annotations are more likely to reflect a particular cultural context [11], which attracts more visitors to the annotated content.

4.3 The Influence of Video Content and Annotators' Behaviours

In this section, we examine whether the annotators' background, annotation behaviour, and video content affected their annotation types. The distribution of annotation types is dominated by the conceptual-level annotations, with the remaining types accounting for a minor proportion. As a result, we focused on the conceptual-level annotations in our analysis. In order to answer RQ2, there are three hypotheses we need to test.

Null Hypothesis 1. *There is no correlation between the annotators' demographic information and the number of different annotation types considered.*

To begin, the diversity of the participants in our study may affect the results, so we used Spearman rank correlation tests (because some variables, such as education level, are ordinal variables) to determine the correlation between annotator demographics and the number of different annotation types. We observed all p values > 0.05, implying no correlation between demographic variables and the number of annotation types. As a result, the Null Hypotheses 1 cannot be rejected, and we can thus ignore the relationship between annotator demographics and experimental results in our data analysis.

Null Hypothesis 2. *Types of annotations are not correlated to annotators' annotation behaviours.*

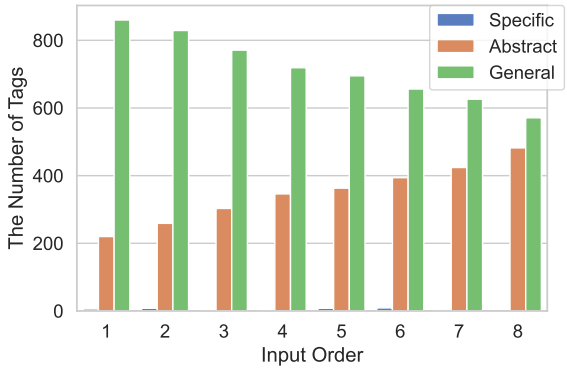

Figure 3: Distribution of the annotations over annotators' input order.

Annotators were required to enter for each video annotations in order. As illustrated in Figure 3, we can see that the number of abstract annotations increases as the input order increases. On the contrary, the number of general annotations decreases as the input order increases. To determine whether there is a correlation between the annotation types and the order of the input annotations, we first convert all of the annotator input types to dummy variables for each annotation. For instance, when an annotation is abstract/Who (e.g., "good health"), the value of the abstract/Who

variable is 1 and the value of all other variables is 0. Because these dummy variables are dichotomous categorical variables and the order of the input annotations is ordinal, the rank biserial test [7] is used to test whether there is a correlation between them.

We can observe some statistically significant correlations between annotation types and input order. First, there is a statistically significant positive correlation between input order and abstract/Who (e.g., "unclean water"), $r(8563) = .069 p < .0005$, input order and abstract/What (e.g., "save water"), $r(8563) = .134, p < .0005$. This indicates that as annotators enter more annotations, the likelihood of the abstract annotation increases. On the other hand, there is a statistically significant negative correlation between input order and general/Who (e.g., "man"), $r(8563) = -.078, p < .0005$, input order and general/What (e.g., "Washing"), $r(8563) = -.114, p < .0005$. This indicates that the likelihood that the annotation is general decreases as the input order increases. Therefore, the Null Hypotheses 2 is rejected, and annotation types are affected by the order that annotators input them.

Annotators' actions, such as the order in which they provide annotations, and how long it takes them to finish to annotate, are tracked by our logger system. As a result, it is possible to determine if the annotators' annotation behaviour correlates with the output annotations. We observe a negative correlation between the time taken by annotators to complete the annotations and general/Who (e.g., "soap"), $r_s(8563) = -.089, p = .003$ with a Spearman's rank-order correlation test due to working duration on a continuous scale and the non-normally distributed data. In addition, under the same test, self-perceived confidence and order of videos are not correlated with annotation types, with $p > 0.05$.

Null Hypothesis 3. *The video variables (length/duration, depicted region and income level) have no effect on the identified annotation types.*

Previous work analysed annotations primarily from the perspective of annotators, comparing whether experts and crowdsourcing annotators produced distinct annotation types and whether familiarity with video content has an impact on annotation types [12], but they lacked an analysis of how different variables within videos of the same type correlate to annotation types. We employed a 4*4 factorial design (four different regions and four different income levels) to pick the videos for our experiments, all of which featured the same action: hand-washing.

We investigated Null Hypotheses 3 based on the region where the families depicted in the video live and their income levels. It is worth mentioning that participating subjects do not directly observe these variables. An aligned ranks transformation ANOVA (ART ANOVA) [37] was conducted to examine the effects of the region, the income level and their interaction effect on the number of different annotation types provided by annotators because of the non-normally distributed data. The ART ANOVA test is a non-parametric test that allows for multiple independent variables and interactions measures. For general/Who annotations (e.g., "hand") that have the largest number of annotations, the main effect for region is statistically significant, $F(3, 1104) = 4.802$, $p = 0.003$ partial $\eta^2 = 0.013$. The main effect for income level ($p = 0.837$) and the interaction effect between region and income level ($p = 0.968$) on the number of general/Who are not statistically significant. A post-hoc pairwise comparison with Tukey adjustment was run and showed a statistically significant Estimated Marginal Means (EMMs) difference between Asia (EMMs=620) and the Americas (EMMs=522), $p = 0.002$, between Asia and Africa (EMMs=542), $p = 0.0214$. This post-hoc analysis shows that annotators prefer to provide more general/Who annotations to the video taken in Asia than in the Americas and Africa.

Additionally, the interaction effect of region and income level on the total number of annotations at the general/What level (e.g., "scrubbing") was not statistically significant, $p = 0.627$. However, the main effect for region at the general/What level was statistically significant, $F(3, 1104) = 3.904$, $p = 0.008$, partial $\eta 2 = 0.01$, and the main effect for income level was also statistically significant, $F(3, 1104) = 4.049$, $p = 0.007$, partial $\eta^2 = 0.011$. The post-hoc analysis for the region revealed that the number of general/What annotations of video shot in Europe (EMMs=608) is statistically higher than in Asia (EMMs=521) and Africa (EMMs=540), $p = 0.008$ and $p = 0.062$ respectively. After conducting a post hoc analysis for income level, it was found that households with low incomes (EMMs=505) received significantly fewer general/What annotations than households with higher incomes (EMMs=591), $p = 0.009$.

Abstract/What (e.g., "poor") is the third most frequently used annotation in our study. The ART ANOVA test revealed that neither the main effect of income level nor the interaction effect between region and income level affected the number of annotations in abstract/What (all p-values were greater than 0.05). However, the region had a statistically significant effect on the number of abstract/What annotations , $F(3, 1104) = 2.926$, $p = 0.033$, and partial $\eta 2 = 0.008$. A post-hoc pairwise comparison with Tukey adjustment demonstrated a statistically significant difference between Asia (EMMs=526) and the Americas (EMMs=600), $p = 0.033$. Abstract/Who (e.g., "good habit") is the next most used annotation, and it was determined that the video's region and income level variables did not affect the number of annotations, with p-values of either main effects or interaction effect from the ART ANOVA test were greater than 0.05.

Finally, we tested whether the length of the video has an effect on the annotation type. Due to the non-normally distributed data and the presence of outliers, Spearman's rank-order correlation tests were performed. Except general/What (e.g., "wash hands"), the relationships between video length and the number of other annotation types were not statistically significant, with all $p > 0.05$. The correlation between video length and the number of general/What annotations (e.g., "using sanitiser") is significant at the 0.05 level ($p = 0.013$) with a minimal correlation coefficient (-0.074).

Discussion. To address RQ2, we analysed the data to determine which factors correlate to annotation types using three hypotheses. In the first hypothesis, we established no correlation between participants' demographic data and the type of annotations they produce. Following that, by testing the second hypothesis, we discovered that the type of annotations made by annotators is influenced by their annotation routines. The time spent by crowdsourcing annotators on the annotation task was negatively correlated with the number of general/who (e.g., "bucket"), indicating that annotators reduce the general description of objects in the video during lengthy tasks.

Additionally, when crowdsourcing annotators provided annotations sequentially, the order of the annotations had a negative statistical correlation with the number of general annotations and a positive statistical correlation with the number of abstract annotations. This means that the first few annotations provided by annotators are typically more general descriptions of objects and actions, whereas the last few annotations are typically abstract and subjective. This finding is similar to [28]. While their definition of abstract/concrete is slightly different from the definition of abstract/general used here, the point remains the same - the later annotations contain more abstract annotations (i.e., more inferences). This provides insight into future crowdsourcing annotation systems' interface design. Specifically, we can use the interface design to influence the order in which crowdsourcing annotators input their content annotations, thereby enriching or controlling the annotation types. The study by Estrada et al. suggest a way to encourage crowdsourcing annotators to provide multiple types of annotations by deploying explicit annotation goals and guidance on interface [12]. Further, we can investigate whether annotation types can be influenced when different incentive mechanisms are used, such as different types of labels offering various incentives.

The data analysis for the third hypothesis revealed that the video content correlates to the number of certain annotation types. First, there is a weak negative correlation between the length of the video and the number of general/What annotations (e.g., "scrubbing hands"). The number of general descriptions of actions in the video decreases slightly as the video length increases. Besides, the number of videos from various SES influences the number of distinct annotation types. High-income households get more general/What annotations , possibly because they wash their hands in more steps such as using soap and hand sanitizer than low-income households, resulting in a more significant number of action description annotations. American households' videos received significantly more abstract annotations than Asian households' videos. The same video content – the typical daily activity of hand washing – but with main characters from disparate SES contexts, resulted in the generation of disparate types of annotations. These variations in annotation types should raise red flags, as they could lead to biased content annotations and, consequently, biased recommendation and search results. For instance, videos from some regions are less likely to appear in search results because they receive fewer annotation types. Guidance with specific instructions and reward incentives may be viable options for resolving the problem.

4.4 Socioeconomic Status Stereotypes

Sentiment analysis. Crowdsourced annotators may be unconsciously projecting their stereotypes in the annotations they provide as annotations created by users can mirror how they describe and view digital resources [23]. Video search engines optimised with these annotations may map these stereotypes [29]. Therefore, in this section, we will address RQ3, by investigating the presence of stereotypes in subjective annotations, also known as abstract annotations. Although a few general annotations might provide some information regarding the stereotypical beliefs of an annotator (i.e., the mention of person, man or woman), it is very hard to form any conclusions without doing a microscopic analysis to all the factors

Negative	Neutral	Positive
263 (9.42%)	1047 (37.51%)	1481 (53.06%)

Table 6: Distribution of abstract annotation sentiment.

that have driven separately each annotators' response (i.e., how long has the video focused on the depicted person). Thus, in this study, we have chosen to instead focus on subjective annotations and carry out a sentiment analysis on those annotations. We used VADER [20], an automated sentiment analysis tool. To measure sentiment, VADER calculates a normalised, weighted composite score between -1 and 1. We set standardised thresholds to classify annotations: a composite score between -0.05 and 0.05 indicates a neutral sentiment; a composite score greater than 0.05 indicates a positive sentiment; a composite score less than -0.05 indicates a negative sentiment. This method of sentiment classification has been widely used [20, 25]. Table 6 shows the number of annotations under the distinct segments of abstract annotations. We found that most of the subjective annotations are positive (e.g., "good habits", "nice video", "living healthy"), accounting for 53.06% of all abstract annotations, but negative annotations (e.g., "dirty", "unhealthy", "bad wash") are still present, accounting for 9.42% of the overall number of abstract annotations.

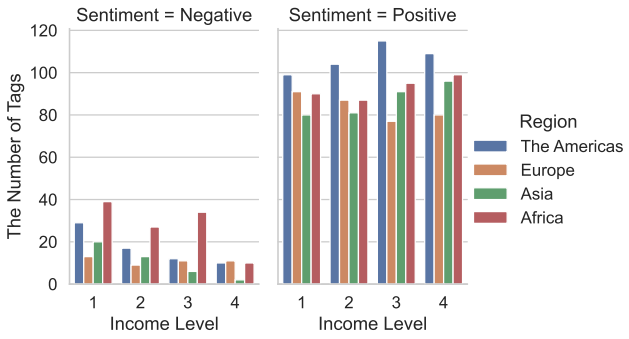

Figure 4: Distribution of different segments of abstract annotations over diverse SES.

We are interested in understanding if and how annotators' stereotypical beliefs are reflected when annotating SES-diverse content. Figure 4 depicts the distribution of negative and positive abstract annotations at various SESs and regions, with values ranging from 1 to 4 representing the different income levels of the households presented in the videos based on Dollar Street numbers, divided into quartiles, 1 representing the lowest 25% and so on. In aggregate, the number of negative annotations received by households with the highest income levels (33) is significantly less than the number of negative annotations received by households with the lowest income levels (101). Households in Africa (110) also received a greater number of negative annotations than in other regions, with 41, 44, and 68 negative annotations received in Asia, Europe, and the Americas, respectively. While the number of positive annotations received by households was similar across income

levels (mean=370.25, SD=10.96), the number of positive annotations received in different regions remained significantly different (mean=370.25, SD=35.2). For example, the number of positive annotations received by families in the Americas (427) is considerably greater than the number of positive annotations received by European households (335). Additionally, as illustrated in Figure 4, there is an interaction effect between the various SESs. For example, high-income African households received fewer negative labels (10) than other African households (all greater than 27), and the gap between the number of negative annotations received by African households and other regions was narrowed.

The selection of videos in our experiment was a standard 4*4 factorial design, and interaction effects were observed from the above analysis of results. Therefore, to answer RQ3, due to the presence of non-normally distributed data and outliers, we performed an ART ANOVA test to analyse whether the different regions, income levels and their interaction effects impacted the VADER compound values, which are unidimensional measures of sentiment. To begin, the main effects for the region and income level on the VADER compound scores were statistically significant, $F(3, 2775) = 3.653$, $p = 0.012$, partial $\eta 2 = 0.004$ for region, and $F(3, 2775) = 7.205$, $p < 0.0005$, partial $\eta^2 = 0.008$ for income level. The interaction effect of region and income level on the VADER compound scores was also statistically significant, $F(9, 2775) = 2.788$, $p = 0.003$, partial $\eta 2 = 0.009$. Spearman's rank-order correlation test results indicate that demographic variables and the VADER compound values are uncorrelated.

The above results show that bias exists when crowdsourcing annotators annotate SES-diverse video content and is influenced by the video's region, income level, and the interaction effects between them in the video. We used Tukey adjustment to perform post-hoc pairwise comparisons on these three factors to determine how these factors specifically influenced the annotators biases. From a regional perspective, the VADER compound scores for videos shot in Africa (EMMs=1328) is statistically lower than in Asia (EMMs=1455) , $p = 0.017$. This means that videos filmed in Africa are more likely to receive negative judgments than those filmed in Asia. The post-hoc test from the perspective of household income level found that statistically significant differences in the VADER compound scores existed between the highest (EMMs=1496) and lowest (EMMs=1299) income levels, $p < 0.0005$. This result implies that videos from low-income people are more likely to be negatively annotated, while videos from high-income people are more likely to be positively annotated. Post hoc tests of interaction effects are much more complex due to having 120 possible combinations. Apart from the stark contrast between lowest-income households in Africa and highest-income households in other regions, there are two combinations worth discussing. The first one is that there is still a significant difference in the VADER compound scores of second higher-income African households (EMMs=1277) than highest-income Asian households (EMMs=1651), $p = 0.001$. In addition, highest-income Asian families receive significantly more positive annotations than the lowest-income households in the Americas (EMMs=1393), $p = 0.01$.

We analyzed annotation behavior to see if individual annotation preferences/biases existed and found no cases where certain annotators preferred to use negative annotations. Also, we performed some Spearman's rank correlation tests and discovered no relationship between annotator demographics and the number of positive and the number of negative annotations they provide, with all p values > 0.05.

Lower class	Lower-Middle class	Upper-Middle class	Upper class
0.305	0.348	0.200	0.176

Table 7: Cohen's k agreement rates between ground truth and annotators' inferred answers for region question across videos with the diverse income level.

Africa	Asia	Europe	The Americas
0.095	0.134	0.100	0.229

Table 8: Cohen's k agreement rates between ground truth and annotators' inferred answers for income level question across videos with the diverse region.

Post-annotation questions. After completing the annotation task, annotators were asked to answer four follow up questions for each video they annotated (see Section 3.1). The first two questions required annotators to estimate the geographic location and economic status of the households depicted in the videos. Due to the fact that annotators could not directly observe these two variables, they could only respond to questions based on their perceptions of the information displayed in the video. Cohen's Kappa was used in our study as a measure of agreement between ground truth data and annotators' inference for the region and income level of family in the video. According to Cohen's $k = 0.257$, there is only a fair agreement between annotators' guesses and actual answers to the first question about where the family lives. Nonetheless, we discover an intriguing phenomenon. Table 7 shows that low-income families have a higher kappa value than high-income families. When running agreement tests for region question across videos with the diverse income level, Lower-income families ($k = 0.305$) and Lower-Middle income families ($k = 0.348$) have higher kappa values than Upper-Middle income families ($k = 0.2$) and Upper income families ($k = 0.176$). This result implies that annotators can better guess the region when confronted with a low-income family in the video, but are less accurate if the video is from a high-income region. Compared to the first question, the agreement between annotators' guessed and true answers for income levels is even lower, with Cohen's $k = 0.139$. Cohen's k agreement rates between annotator guesses and actual income levels for videos from various regions are displayed in Table 8. In the footage from Africa, we found that annotators obtained the lowest kappa rate ($k = 0.095$) for income levels, while in the video from the Americas, annotators obtained the highest kappa rate ($k = 0.229$).

The final two follow-up questions asked annotators to rate the video's suitability for inclusion in Google search results and public service announcements on a five-point Likert scale ranging from '1: Not at all' to '5: Totally'. Table 9 and Table 10 present the results for those two questions. We discover that as the income level of the households featured in the videos increases, these videos are more

	Africa	Asia	Europe	The Americas	All
Lower class	2.91	2.90	3.33	3.01	3.04
Lower-Middle class	2.97	3.01	3.69	3.29	3.24
Upper-Middle class	3.01	3.53	3.73	3.47	3.44
Upper class	3.29	3.56	3.79	3.56	3.55
All	3.05	3.25	3.63	3.33	3.32

Table 9: Annotators' perspective on the suitability of the video to appear in Web search results (average on a 1-5 scale).

	Africa	Asia	Europe	The Americas	All
Lower class	2.86	2.83	3.03	2.90	2.90
Lower-Middle class	3.00	2.99	3.53	3.36	3.22
Upper-Middle class	3.09	3.43	3.60	3.51	3.41
Upper class	3.49	3.70	3.57	3.49	3.56
All	3.11	3.24	3.43	3.31	3.27

Table 10: Annotators' perspective on the suitability of the video to appear in a public service announcement (average on a 1-5 scale).

likely to be deemed appropriate by annotators. Additionally, videos from various regions receive varying ratings. For instance, African videos receive lower average ratings than videos from other regions. Two ART ANOVA tests showed statistically significant effects of region and income level on the video on annotators' answers to these two questions. The main effect of the region on the public announcement had a p-value of 0.02 and all other $p < 0.0005$.

For each main effect, all pairwise comparisons with Tukey adjustment were conducted. The post-hoc test for the region on the suitability of the video to appear in google search results revealed that videos from Africa (EMMs=488) received statistically significantly lower EMMs than videos from Europe (EMMs=658, $p < 0.0005$) and the Americas (EMMs=568, $p = 0.017$), while videos from Europe received statistically significantly more EMMs than videos from Asia (EMMs=529, $p < 0.0005$) and the Americas ($p = 0.005$). In terms of income, videos from low-income households (EMMs=491) received statistically significantly lower ratings for search results than videos from upper-middle-income households (EMMs=573, $p = 0.014$) and upper-income households (EMMs=635, $p < 0.0005$). Additionally, videos from lower-middle-income households (EMMs=491) received statistically significantly lower ratings than those from upper-income households ($p = 0.005$). Moving on to the question of the video's suitability for inclusion in a public service announcement, a post-hoc analysis for the region revealed that videos from Africa (EMMs=525) received statistically significantly fewer EMMs than videos from Europe (EMMs=607, $p = 0.016$). As demonstrated by the post-hoc test for income level, there were statistically significant differences in ratings between videos from low-income (EMMs=475) and upper-middle-income households (EMMs=582), $p < 0.0005$, between videos from low-income and upper-income households (EMMs=645), $p < 0.0005$, between videos from lower-middle-income households (EMMs=582) and upper-income households, $p = 0.001$. Moreover, the interaction effect between the region and income level did not significantly affect the suitability

	Africa	Asia	Europe	The Americas
GRC	2.83	3.11	3.74	3.4
GRW	3.33	3.34	3.6	3.23
All	3.05	3.25	3.63	3.33

(a)

	Africa	Asia	Europe	The Americas
GRC	2.75	3.09	3.52	3.44
GRW	3.57	3.32	3.4	3.12
All	3.11	3.24	3.43	3.31

(b)

Table 11: Annotators who guess region correctly (GRC) and Annotators who guess region wrongly (GRW). Results show their perspective (average on a 1-5 scale) on the suitability of the video from four regions to appear in (a) Web search results and (b) a public service announcement.

	Lower class	Lower-Middle class	Upper-Middle class	Upper class
GIC	2.77	3.16	3.63	4.06
GIW	3.18	3.52	3.69	3.4
All	3.04	3.24	3.44	3.55

(a)

	Lower class	Lower-Middle class	Upper-Middle class	Upper class
GIC	2.59	3.23	3.71	3.98
GIW	3.27	3.21	3.27	3.47
All	2.9	3.22	3.41	3.56

(b)

Table 12: Annotators who guess income level correctly (GIC) and Annotators who guess income level wrongly (GIW). Results show their perspective (average on a 1-5 scale) on the suitability of the video from four income levels to appear in (a) Web search results and (b) a public service announcement.

for inclusion in Google search results and public announcements, with p-values of 0.85 and 0.48, respectively.

In the first two questions, we discovered that some annotators may have misclassified some videos region and income level. Table 11 and Table 12 show rating scores grouped by annotators who guessed region and income level correctly and incorrectly. We observe that annotators who correctly guessed the video region gave lower scores than those who incorrectly guessed the region for videos from Africa and Asia. In contrast, for videos from Europe and the Americas, annotators who correctly guessed the video region tended to give higher scores than annotators who incorrectly guessed the region. We used Mann-Whitney U tests to determine if there were significant differences. For videos from African households, annotators who correctly guessed the region (mean rank = 125.16, 120.05) would provide statistically significantly lower scores than incorrectly guessed annotators (mean rank = 160.36, 166.98) on the questions about search results or public announcement suitability, $U = 7215, z = -3.74, p < 0.0005$ and

$U = 6407.5, z = -4.925, p < 0.0005$, respectively. Furthermore, this significant difference was also present in videos from the American continent. On the question about public announcements, annotators who correctly guessed the region of the video (mean rank = 149.28) scored video suitability statistically significantly higher than annotators who incorrectly guessed the region (mean rank = 127.13),$U = 10864, z = 2.311, p = 0.021$.

Furthermore, we discovered that annotators who correctly guessed the income level for videos from the lowest income level tended to give lower suitability scores than those who incorrectly guessed the income level. Annotators who correctly guessed the income level for videos from the highest income level bracket tended to give higher scores than those who incorrectly guessed the income level. The Mann-Whitney U tests showed that the difference described above was significant in questions about search results and public announcements, except for videos from the lower-middle class. For the videos from the lower-middle class, correctly guessing annotators (mean rank = 121.77, 120.91) provided statistically significantly lower scores than incorrectly guessing annotators (mean rank = 162.75, 163.76), $U = 6880.5, z = -4.361, p < 0.0005$ and $U = 6751, z = -4.517, p < 0.0005$, respectively. Correctly guessing annotators (mean rank = 154.08, 159.6) provided significantly higher scores than incorrectly guessing annotators (mean rank = 134.17, 131.6) for videos from the upper-middle class, $U = 9708, z = 2, p = 0.045$ and $U = 10199, z = 2.783, p = 0.005$, respectively. For the videos from the Upper class, correctly guessing annotators (mean rank = 181.69, 171.95) provided statistically significantly higher scores than incorrectly guessing annotators (mean rank = 131.76, 133.83), $U = 7678, z = 4.099, p < 0.0005$ and $U = 7200.5, z = 3.126, p = 0.002$, respectively.

Discussion. We observed evidence of bias based on region and income level when we conducted a sentiment analysis of abstract annotations generated by crowdsourced annotators (RQ3). First, videos depicting low-income households were more likely to receive negative annotations, whereas videos with higher-income families received more positive annotations. The highest-income households received significantly fewer negative annotations than the lowest-income households. Second, negative annotations were more prevalent for videos shot in Africa than in Asia. We also observed evidence of bias based on region and income level in our analysis of the results from the post-task questions (RQ4). To begin, annotators could not accurately guess the region and income level of the households in the video. This suggests that annotators' choice of annotations is solely based on the information available in the video. Such judgments can lead to implicit bias rather than explicit bias. When annotators are confronted with low-income households in the video, they make more accurate guesses about their region than those with high-income. We noticed many low-income households had videos where they washed their hands outdoors while high-income households mainly were inside so that the region may have been more apparent. Compared to other regions, annotators were less accurate in guessing families' income levels for African videos, possibly due to subconscious stereotypes. By contrast, annotators were more accurate in guessing income levels of households in the Americas because our study participants were from the United States and cultural exposure likely plays a role. In addition, annotators' perspectives towards videos varied according to region and income level. Annotators deemed videos from higher income groups (Europe and the Americas) more appropriate for inclusion in search results and public service announcements as compared to videos from Africa and lower-income countries. This can be explained by the fact that high-income households might have more resources to follow hygiene guidelines, resulting in higher ratings. Also, as there is no "ground truth" annotation for each video in the final two questions, the higher ratings for high-income groups may not accurately reflect the level of bias of the annotators. Nevertheless, we can observe this bias by comparing the ratings of annotators who guessed the correct region and income with those who did not. In particular, annotators who failed to guess the video location gave higher suitability scores to videos from Africa and low-income levels, and lower scores when they guessed correctly. In addition, the phenomenon of videos receiving different ratings across regions and income levels can also be explained by the uneven representation of SES levels in search engines and public service announcements.

There is a risk that the over-representation of Western and high-income populations and regions in crowd-generated annotations may perpetuate stereotypes about other countries or specific social groups in those countries. De Vries et al. employed six public object detection cloud services to determine household items in the Dollar Street dataset and showed that these cloud services are biased towards data from different income groups and geographic locations [9]. High-income households benefit significantly more from these cloud services in terms of object classification accuracy than low-income families do. It also exists across regions, with, e.g., the Amazon Rekognition system being more accurate for household objects photographed in the United States than for objects photographed in other countries. These findings are consistent with the biases discovered in our study, implying that the biases observed in publicly available cloud services may be the result of pre-trained algorithms using similar human-generated annotations. The annotations' biases are inherited by the algorithms. Therefore, the first step towards addressing these biases in public cloud services is to minimise their presence during the annotation phase.

5 CONCLUSION

To answer our research questions, we conducted a controlled user study using crowdsourcing, in which US participants are asked to annotate SES-diverse content. Our findings indicate the following. RQ1: In line with previous research, we discovered that general/Who (e.g., "plate"), general/What (e.g., "use soap") and abstract/What (e.g., "neat") were the most frequently used annotation types. We discuss the application of these various types of annotations and demonstrate several strategies for encouraging users to provide a variety of annotations. RQ2: The order in which annotators enter annotations influences the type of annotations. We observe a statistically significant negative correlation between the order of annotations and the number of general descriptive annotations and a statistically significant positive correlation with the number of abstract annotations and subjectivity. Second, the number of videos from particular SESs correlated with the distinct annotation types. High-income households received more

descriptive annotations for actions, and videos from households in the Americas received significantly more subjective annotations than videos from Asian households. RQ3: We observe regional and income-level biases when performing sentiment analysis on crowdsourced annotator-generated abstract annotations. Negative annotations are used more frequently for African videos than those from Asia. RQ4: We discovered bias in the analysis of the follow-up questions based on region and income level. Annotators determined that videos from higher-income groups (i.e., Europe and the Americas) were more appropriate for search results and public service announcements. When designing data annotation tasks, it is important to consider these findings to ensure a more diverse and less-biased set of annotations. For example, it would be useful not only to conceal metadata such as content location, but also to collect information about annotators' awareness of such metadata to then take this into account during annotation post-processing.

One limitation of our study is that the results are based on annotators from the United States. In future research, we can take the location of the annotators into account to conduct a more comprehensive investigation. For example, we found that annotators considered videos from the Americas and Europe more appropriate for inclusion in search results or public service announcements than videos from Africa. However, it is not entirely clear how much of this was biased, so it would be interesting to compare them with annotations obtained from annotators based in other locations.

Acknowledgments. This work is partially supported by the the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT), an ARC Discovery Project (Grant No. DP190102141), and the ARC Training Centre for Information Resilience (Grant No. IC200100022).

REFERENCES

[1] Morgan Ames and Mor Naaman. 2007. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 971–980.
[2] Linda H Armitage and Peter GB Enser. 1997. Analysis of user need in image archives. *Journal of information science* 23, 4 (1997), 287–299.
[3] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. 2021. To "See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–31.
[4] Pınar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2019. Social b (eye) as: Human and machine descriptions of people images. In *Proceedings of ICWSM*, Vol. 13. 583–591.
[5] Sebastian Chan. 2007. Tagging and Searching–Serendipity and museum collection databases. In *Museums and the Web*, Vol. 2007. 87–99.
[6] Alessandro Checco, Jo Bates, and Gianluca Demartini. 2018. All that glitters is gold—An attack scheme on gold questions in crowdsourcing. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.
[7] Edward E Cureton. 1956. Rank-biserial correlation. *Psychometrika* 21, 3 (1956), 287–290.
[8] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2014. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491* (2014).
[9] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 52–59.
[10] Federica Durante and Susan T Fiske. 2017. How social-class stereotypes maintain inequality. *Current opinion in psychology* 18 (2017), 43–48.
[11] Irene Eleta and Jennifer Golbeck. 2012. A study of multilingual social tagging of art images: cultural bridges and diversity. In *ACM 2012 Conference on Computer Supported Cooperative Work*. 695–704.
[12] Liliana Melgar Estrada, Michiel Hildebrand, Victor de Boer, and Jacco van Ossenbruggen. 2017. Time-based tags for fiction movies: comparing experts to novices using a video labeling game. *Journal of the Association for Information Science and Technology* 68, 2 (2017), 348–364.
[13] Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu. 2012. Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing* 22, 1 (2012), 363–376.
[14] Gapminder. 2021. Dollar Street - photos as data to kill country stereotypes. https://www.gapminder.org/dollar-street
[15] Riste Gligorov, Michiel Hildebrand, Jacco Van Ossenbruggen, Guus Schreiber, and Lora Aroyo. 2011. On the role of user-generated metadata in audio visual collections. In *The sixth international conference on Knowledge capture*. 145–152.
[16] Scott Golder and Bernardo A Huberman. 2005. The structure of collaborative tagging systems. *arXiv preprint cs/0508082* (2005).
[17] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. The impact of task abandonment in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* (2019).
[18] L Hollink, A Th Schreiber, Bob J Wielinga, and M Worring. 2004. Classification of user image descriptions. *International Journal of Human-Computer Studies* 61, 5 (2004), 601–626.
[19] Anne-Stine Ruud Husevåg. 2017. Categorization of Known-Item Search Terms in a TV Archive. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 321–324.
[20] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.
[21] F Jahanbakhsh, J Cranshaw, S Counts, W S Lasecki, and K Inkpen. 2020. An Experimental Study of Bias in Platform Worker Ratings: The Role of Performance Quality and Gender. In *The 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[22] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *CHI Conference on Human Factors in Computing Systems*. 3819–3828.
[23] Judith L Klavans, Rebecca LaPlante, and Jennifer Golbeck. 2014. Subject matter categorization of tags applied to digital images from art museums. *Journal of the Association for Information Science and Technology* 65, 1 (2014), 3–12.
[24] Michael W Kraus, Jun Won Park, and Jacinth JX Tan. 2017. Signs of social class: The experience of economic inequality in everyday life. *Perspectives on Psychological Science* 12, 3 (2017), 422–435.
[25] Fritz Lekschas, Spyridon Ampanavos, Pao Siangliulue, Hanspeter Pfister, and Krzysztof Z Gajos. 2021. Ask Me or Tell Me? Enhancing the Effectiveness of Crowdsourced Design Feedback. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.
[26] Jesse S Michel, Sadie K O'Neill, Paige Hartman, and Anna Lorys. 2018. Amazon's Mechanical Turk as a viable source for organizational and occupational health research. *Occupational Health Science* 2, 1 (2018), 83–98.
[27] Jahna Otterbacher. 2015. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1955–1964.
[28] Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. 2019. How do we talk about other people? group (un) fairness in natural language image descriptions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 106–114.
[29] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *The 2017 CHI conference on human factors in computing systems*. 6620–6631.
[30] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating user perception of gender bias in image search: the role of sexism. In *The ACM SIGIR conference on research & development in information retrieval*. 933–936.
[31] Erwin Panofsky. 2018. *Studies in iconology: humanistic themes in the art of the Renaissance*. Routledge.
[32] Ludovic Righetti, Raj Madhavan, and Raja Chatila. 2019. Unintended consequences of biased robotic and artificial intelligence systems [ethical, legal, and societal issues]. *IEEE Robotics & Automation Magazine* 26, 3 (2019), 11–13.
[33] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. Association for Computing Machinery, New York, NY, USA, 2863–2872.
[34] Sara Shatford. 1986. Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly* 6, 3 (1986), 39–62.
[35] Veslemøy Søbak and Nils Pharo. 2017. Decentralized subject indexing of television programs: The effects of using a semicontrolled indexing language. *Journal of the Association for Information Science and Technology* 68, 3 (2017), 739–749.
[36] Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. 2012. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys (CSUR)* 44, 4 (2012), 1–24.
[37] J O Wobbrock, L Findlater, D Gergle, and J J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *SIGCHI conference on human factors in computing systems*. 143–146.