Figure 1: Examples of potentially disturbing content (both are material used in the card sorting activity).

Saudis' Conceptualisation of Disturbing Content on Social Media

Sara Albakry
ssbakry@uqu.ed.sa
Umm Al-Qura University
Makkah, Saudi Arabia
University of Edinburgh
Edinburgh, UK

Aljawharah Alabdullatif
alalabdullatif@ksu.edu.sa
King Saud University
Riyadh, Saudi Arabia
University of Edinburgh
Edinburgh, UK

Kami Vaniea
Maria Wolters
kvaniea@inf.ed.ac.uk
maria.wolters@ed.ac.uk
University of Edinburgh
Edinburgh, UK

ABSTRACT

Frequent exposure to disturbing content on social media such as posts, sharing, accident news, or even photos of puppies could adversely impact users' online experience or well-being. Several protection mechanisms exist to provide users with control over content feeding into their personal spaces; such as sensitive and "show less often" markers. Better understanding of users' conceptions of disturbing content is a prerequisite to providing users with their desired level of privacy. As a first step, we designed a protocol combining two requirement elicitation techniques: affinity diagram and card sorting. In this paper, we report our protocol and reflect on a pilot with two transnational Saudi women groups. Finally, we suggest possible future research directions to further improve our understanding of Arab users' needs and practices in the context of personal space maintenance.

KEYWORDS

Definition *Disturbing content* is defined in this context as content disruptive to the social media experience or something users would prefer not to see in their feed. This definition emphasises our interest in learning personal needs instead of societal needs.

Research Question How do transnational Saudi women conceptualise disturbing content on social media?

Methodology Question Does using a combination of minimally structured affinity diagrams followed by more structured card sorting activities overcome issues of social desirability and unformed mental models?
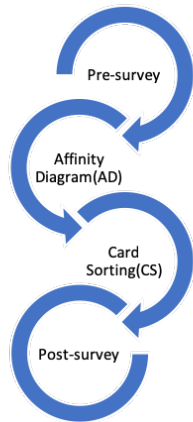

Figure 2: Disturbing Content Elicitation Protocol

INTRODUCTION

Social networks allow sharing of information between people at impressive scale and speed; doing so helps people be socially connected which in turn provides the potential to improve their general well-being. However, not all people react to social networks or the content they provide in the same way [1,3]. Content that one person views as positive and empowering, may be seen by someone else as stress inducing. An example might be phobias or cultural elements that are seen differently in different parts of the world. Dogs, for example, can be seen as anything from cherished family members (pets), memes (ICanHas.Cheezburger.com), impure (some Muslims' views), or physically dangerous (urban wild dogs). Unlike illegal or physically violent content, these types of content are very challenging for social networks to automatically handle because they are problematic for the individual, rather than for the larger collective network or culture.

In this work, we narrow our focus to people's boundaries on social media and how violations of those boundaries ultimately impact well-being. Karr-Wisniewski et al. identified ten different boundaries people use to regulate their privacy, here we focus on the territory boundary (inward-facing), defined as "regulating incoming content for personal consumption" [2]. We find the territory boundary particularly interesting to study, because it involves managing the accumulation of stress inducing events which individually cause minimal impact but when accumulated over time can materially impact well-being of an individual. However, a lack of connectedness can also be a cause of stress and lead to missing important information. Therefore, users need strong individual-level controls that allow them to manage not just who can see content they generate, but also the content they consume. Measuring the effectiveness of these controls can be challenging because as a community we currently have a limited understanding of the types of content that different individuals consider disturbing or disruptive. We also lack an understanding of how widely the perceptions of content varies between individuals, cultures, and groups.

One of the challenges for this type of research is *elicitation*. The type of information we are looking to learn about is in itself a sensitive topic and any study design eliciting it needs to take into account that sensitivity. It is also on a topic that is susceptible to social normative pressures and the privacy paradox [5], both of which bias people towards giving answers that are more socially acceptable and not necessarily grounded in their actual behaviour or opinions. The topic is also one that few people have devoted time to think about, which means that participants are likely to have underdeveloped mental models. Therefore, to study such a phenomena we need to create a study methodology that both creates a safe space for sharing sensitive opinions while also providing support for participants to develop their own understanding over the course of the study. To accomplish both goals we decided to use a combination of an affinity diagram which involves both individual brainstorming and group sorting, and a card sorting activity, which might help participants broaden their understanding of the problem space.

In this paper, we present our chosen methodology and discuss the results of a pilot with two groups of Saudi women. In particular, we discuss how our methodology decisions impacted our ability to collect the types of data we are interested in.

Table 1: Participants' Demographics

P#	Age Group	Education
P1	31-35	Postgraduate
P2	26-30	College graduate
P3	26-30	Postgraduate
P4	26-30	Postgraduate
P5	26-30	Postgraduate

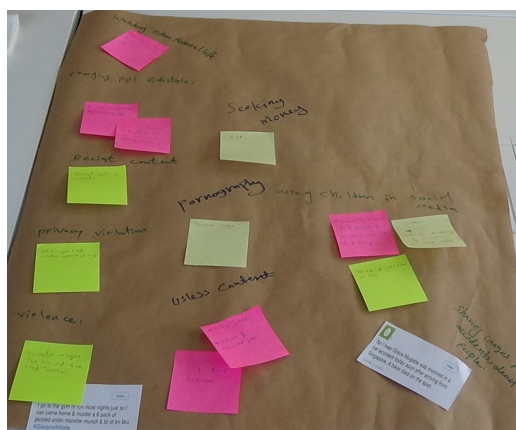**Figure 3: Affinity diagram produced by the first focus group.**

METHOD

Our aim is to understand users' conceptions of disturbing content through the collection of the following information: 1) disturbing content categories from individuals, 2) sources of individual sensitivity and 3) sources of individual or group disagreement and 4) other demographic and contextual information, such as social media usage and their relationship to people in their network. The major challenge here was getting participants to express their actual opinions free from biases such as social normative, or the privacy paradox [5]. The study took place at the University of Edinburgh in the United Kingdom (UK), following the School of Informatics ethics procedure[1].

Recruitment: Participants were recruited by circulating a recruitment message to local Edinburgh individuals and groups of Saudi females via WhatsApp. The message included information about the research purpose, researchers contact information, a link to sign up, and the expected duration (an hour). Participation was voluntarily with no compensation.

Participants: Five female participants responded to the advertisement and joined the study, they all specified "Arabic" as their native language (Table 1). Three participants participated in the first session while the rest participated in the second session.

Study Design: We chose a focus group setting to be able to collect both individuals' and groups' conceptions as well as observe non-verbal cues and group dynamics. To add more structure to the discussion, we combined two techniques: affinity diagram (AD) and card sorting (CS). AD allows participants to express and sort their own ideas, while CS provides concrete examples that are uniform across groups allowing between group comparisons. As the protocol was designed to allow replication across different cultures, English was the main language used verbally and on written instructions. However, participants in the pilot were also encouraged to use code-switching in situations where language barriers interfered with their ability to express their ideas or feelings. Due to the sensitive nature of the topic, we decided not to audio/video record the sessions so as to allow participants to express their ideas more comfortably.

Procedure: The session started with a consent form detailing the content of the study and explicitly asking for permission to take pictures of the resulting AD and CS diagrams. Then participants were asked to fill out a short survey that collected demographic data, information about their social media platform usage and Westin's Privacy Segmentation Index Questionnaire [4]. The main session consisted of first building an AD, and then the CS using two different sorting methods. For the AD, the participants were asked to individually think of and write down three types of social media content that they found disturbing. Next, participants were asked to collectively group and label the types of disturbing content they generated to create an AD (Figure 3). For the CS, a set of cards with mock tweets were given to participants. First, they were asked to classify them into two sets: disturbing and not disturbing. Once finished, they were asked to sort the cards again, but this time using the affinity diagram they previously generated. They were given the option to expand the AD as needed. Finally, participants were asked to fill out a post survey about who they followed in each social media platform and which social media platforms they encounter disturbing content and how often.

Table 2: Final Generated Categories

P#	Categories	Count
P1	Ads, sexual images, videos taking advantage of children to seek help	3
P2	Racist posts or comments, text or image that violates someone's privacy, violent images and the use of the word murder, the use of children in ads	4
P3	Silly food pictures, Large volume of silly non-sense jokes, presenting personal life and details of their relationships with kids and family, people presenters their personal and private life, using social media for showing off, change their lifestyle to be seen in a luxury way	6
P4	Ads, blood or surgical content, any-thing related to wars, kids acting like adults	4
P5	Pornographic images, sex and nudity, Bullying sexiest, racist...etc, pictures of war victims, sex, violence, drugs, political propaganda	7

Table 3: Social Media Platforms Usage and Frequency of Disturbing Content

P#	Twitter	Facebook	Instagram	Snapchat
P1	D:W	R:NA	D:D	D:D
P2	NA:NA	NA:NA	D:D	D:R
P3	NA:NA	R:NA	W:D	D:D
P4	M:M	R:NA	D:W	D:D
P5	D:D	NA:NA	R:NA	D:M

D : Daily   W:Weekly   M:Monthly   R: Rarely
NA: Not Used/Not Applicable

OUTCOME AND LEARNED LESSONS

In this section, we reflect on aspects of the protocol, collected data and procedure. Figure 3 shows the affinity diagram generated by the first focus group.

Impact of word choice "Disturbing" on emerging categories: From the generated categories shown in Table 2, we observed a wide range of participant-created categories, such as illegal (e.g. privacy violation), inappropriate (e.g. discriminatory jokes) and sensitive (e.g. injured bodies) suggesting that "disturbing" as a term was sufficiently inclusive to solicit various sensitivity sources.

English cards triggered cultural-comparison: During the card sorting activity, participants often used the pronouns "we" and "they" to refer to members of their home culture and hosting culture, respectively, possibly caused by the fact that the sample tweets were written in English and the homogeneity of the participants in terms of language and national origin. One group's opinions were split over a card with a picture of a dog. While one classified it as not disturbing and even described it as "cute" another thought it was a disturbing image, saying that "A dog is impure". This difference of opinion lead to a discussion on if this type of content was likely to be posted by people in their own cultures, and if the source changed whether it was seen as disturbing.

Note taking challenges due to no audio recording: Not recording audio created a safe space to share potentially sensitive opinions, but it also made it challenging for the researchers to comprehensively record participant interactions. A potential solution might be to add elements that support note taking, such as clear numbers on the cards which the note taker could reference quickly.

Prompting questions to add grounding: Although the individual brainstorming activity generated a range of disturbing content categories, some of the ideas were more abstract than grounded in past events. The protocol needs a set of prompting questions which could be used to further contextualise what participants mean by these categories. For example, one participant mentioned ads promoting a site for dating Saudi women as disturbing. However, a follow up discussion revealed that it was a web-based ad not a social media ad.

Rethinking the post-survey component: Table 3 shows participants' responses to the post-survey, including how often they have used different social media platforms and encountered disturbing content in each of these platforms. While this is useful information for forming a holistic view of participants' perception of their social media experience, including it in future focus groups is still an open question considering the time it took and the minimal added value.

Focus group atmosphere: The AD activity created two different atmospheres. The debriefing part triggered a range of emotions which varied across shared experiences: sympathy, fear, anger, sadness and helplessness. On the contrary, in the grouping part, participants seemed empowered and expressed enthusiasm while moving content around.

FUTURE WORK

Understanding how different people conceptualise disturbing content is important for multi-national social media organisations like Twitter and Facebook to meet their users' needs, especially as they struggle to manage inappropriate content on their platforms. This research shows through a pilot study how this type of research could be conducted in a privacy sensitive way while still generating meaningful outcomes.

The future work for this project involves looking at a larger sample of multi-national participants from the Arab world and beyond. With the eventual goal of taking the results and converting them in a survey format to better ask questions about the prevalence of concern about the various forms of content identified.

REFERENCES

[1] Jonah A Berger and Eva Buechel. 2012. Facebook therapy? why do people share self-relevant content online? (2012).Available at SSRN: http://ssrn.com/ abstract=2013148 or http://dx.doi.org/10.2139/ssrn.2013148.
[2] Pamela Karr-Wisniewski, David Wilson, and Heather Richter-Lipford. 2011. A new social order: Mechanisms for social-network site boundary regulation. InAmericas Conference on Information Systems, AMCIS.
[3] Ethan Kross, Philippe Verduyn, Emre Demiralp, Jiyoung Park, David Seungjae Lee, Natalie Lin, Holly Shablack, JohnJonides, and Oscar Ybarra. 2013. Facebook use predicts declines in subjective well-being in young adults.PloS one8, 8(2013), e69841.
[4] Ponnurangam Kumaraguru and Lorrie Faith Cranor. 2005.Privacy indexes: a survey of Westin's studies. Carnegie Mello-nUniversity, School of Computer Science, Institute for Software Research International.
[5] Pamela Wisniewski, Jessica Vitak, Xinru Page, Bart Knijnenburg, Yang Wang, and Casey Fiesler. 2017. In Whose BestInterest?: Exploring the Real, Potential, and Imagined Ethical Concerns in Privacy-Focused Agenda. InCompanion of the2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17 Companion). ACM, NewYork, NY, USA, 377–382. https://doi.org/10.1145/3022198.3022660