A Technology Exploration towards Trustable and Safe Use of Social Media for Vulnerable Women based on Islam and Arab Culture

MIRKO FRANCO, University of La Rochelle, France
SALAH A. FALYOUN, Tokyo University of Foreign Studies, Japan
KAREN E. FISHER, University of Washington, USA
OMBRETTA GAGGI, University of Padua, Italy
YACINE GHAMRI-DOUDANE, University of La Rochelle, France
AYAT J. NASHWAN, Yarmouk University, Jordan
CLAUDIO E. PALAZZI, University of Padua, Italy
MOHAMMED SHWAMRA, University of Washington, USA

We live in an era characterized by unprecedented communication possibilities; yet significant parts of the world's population are left behind and at increased vulnerability of harm due to low digital literacy, poverty, war, surveillance, gender and cultural bias, etc. Social media platforms are not perceived as trustable and safe (e.g., privacy-preserving) enough by many users, especially by the most vulnerable ones. In this context, we aim at creating a bridge between vulnerable users and computer science to let researchers know where they should address their efforts to improve these tools. In particular, utilizing the results from ongoing fieldwork with conservative Sunni-Muslims, mostly Syrian war refugees, in Jordan, we identify some crucial features that messaging systems should include, discuss how they are indeed important for any user, and propose some possible technological approaches to implement them. Finally, we also discuss how we have added one these highly ranked features to a custom messaging system, proving its feasibility.

CCS Concepts: • **Software and its engineering → Requirements analysis**; • **Human-centered computing → User interface programming**.

Additional Key Words and Phrases: messaging systems, privacy, social networks

ACM Reference Format:
Mirko Franco, Salah A. Falyoun, Karen E. Fisher, Ombretta Gaggi, Yacine Ghamri-Doudane, Ayat J. Nashwan, Claudio E. Palazzi, and Mohammed Shwamra. 2022. A Technology Exploration towards Trustable and Safe Use of Social Media for Vulnerable Women based on Islam and Arab Culture. In *Conference on Information Technology for Social Good (GoodIT'22), September 7–9, 2022, Limassol, Cyprus.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3524458.3547259

1 INTRODUCTION

People worldwide are being empowered by access and fluency in social media; however, significant pockets are being left behind, their situations exasperated by war, poverty, gender and ethnic bias, cultural norms, surveillance, and hacking. Aimed at creating positive social initiatives that support privacy protection, this study addresses how culturally based technologies codesigned with at-risk people can improve their daily lives and mitigate biases. The primary foci of the study are Syrian and Jordanian females whose lives are governed by conservative Sunni Muslim and Arab cultural norms regarding privacy, identity, communication and mobility.

Arab social media (SM) use—Facebook, YouTube, Instagram, WhatsApp, Tiktok, Snapchat, Telegram, etc.—is exploding with male users privileged by mobile ownership and cultural norms [18, 25]. Social media is a lifeline for Arab females supporting communication, information, livelihoods, and human rights—female empowerment; however, research is severely lacking on women' social media experience and their digital literacy (DL), and online privacy needs due to weak methodology and reporting, specifically regarding field access, trust, and cultural understanding.

The Syrian war, in its 11th year, has externally displaced over 13 million people with most refugees living in Jordan, Turkey, and Lebanon [10, 11]. The shredding of Syria's social fabric means increased household and economic responsibilities for females that exasperate lost years of schooling and opportunity. Of all the victims of the Syrian war, girls and women have suffered most: subject to early marriage, human trafficking, poverty, and gender-based violence and harassment [4]. Jordan - a resource-poor country with limited agriculture land, no oil resources and a scarce water supply, shares borders with Syria, Iraq, Saudi Arabia, and Palestine-Israel. Mostly Sunni Muslim (over 95%), its patriarchal society of over 10M people is predominantly young with 74% under age 30. Decades of hosting war refugees has burdened Jordan's economy and resources. Covid19 saw poverty skyrocket with youth unemployment exceeding 50%, and, for females, increased household responsibilities and gender-based violence (GBV), including torture and honor killings (e.g., [17]). As cultural norm, women do not work in most sectors and quit work after marriage [27, 30] with refugees having steeper work sector restrictions.

Social media is valued for its visual richness, seeing the lives of friends and celebrities, and creative self-expression that can hide women's circumstances and promote sense of normalcy—fueled by followers, likes, comments and shares. Yet, these features conflict with women and girls' lives, risking safety and standing. Females in Jordan are often targeted by scammers and trolls with dis-, mis-, and mal-information, compounding their vulnerability to violating Islamic and cultural norms of identity, mobility, talking with non-family users, posting photos, wealth status, and location [9, 10].

Our work draws on findings from ongoing fieldwork with Sunni Muslims, in Jordan for which one privacy feature was developed. The paper is intended as a bridge between users and the computer science field to advise developers of where and how they should focus their effort. Specifically, we leveraged on ongoing fieldwork with conservative Sunni-Muslims, mostly Syrian war refugees, in Jordan, to identify major key features that should be embedded in messaging systems to make them more trustable and safe. We discuss how they are indeed important for any user and discuss possible technology to actually deploy them. Finally, as a starting point, we present the implementation into a custom messaging system of one of the aforementioned features that has been ranked among the most important ones.

The remainder of this paper is organized as follows. Section 2 presents a review of the related scientific literature. Section 3 describes the motivation behind our work. Section 4 discusses some features for safer and trustable social media and potential technological approaches. Section 5 presents a preliminary proof-of-concept messaging platform that embodies one of the discussed functionalities. Finally, we draw our conclusions in Section 6.

2 RELATED WORK

The concept of trustable and safe use of social network is very broad and includes diverse aspects such as privacy preservation, trustfulness of information, reputation, reliability and more. For instance, Halevy *et al.* [16] have presented a review of the recent progress and the opening problems about keeping online social media platforms and their users safe from malicious activities (e.g., misinformation, deepfakes, multimodal malicious content, coordinated action, etc.). Amon *et al.* [7] have investigated the implications for children's privacy of the sharing of their photos by their parents on online social media, defining design implications to guarantee responsible sharing of children's information on online platforms. Other works [14, 21], just to mention some examples, discuss the design of parental control applications for a safe experience on mobile applications and social networks, considering children and teenagers, to minimize online related risks.

However, social media and messaging platforms are far away from being safe and trustable. For instance, the commonly used messaging applications allow users to send any content to anyone, thus permitting the spreading of personal content without the owner's consent [12]. Oukemeni *et al.* [23] have analyzed some social networks and their characteristics (e.g., architecture, functionalities, encryption, provided services, etc.), focusing on microblogging ones, acknowledging that current privacy mechanisms are not sufficient to protect users and their content. As a first step towards safer and trustable social media systems, Franco *et al.* [12] have proposed some guidelines for developers to build messaging platforms safer than the current ones for sexting thanks to a forwarding control feature for images.

Furthermore, privacy is a term with different meanings depending on the culture, the society, the religion, and the platform itself [23]. For instance, the vision of privacy of Arabic users is influenced by Islamic teachings and culture: maintenance of reputation, and the preservation of respect, modesty and humility, while it is focused on concerns about safety and security in the Western cultures [2]. The different conceptions of privacy impact how people use social media; designers need to consider cultural values and their impact while building online platforms. Indeed, designing for Arabic users is one of the greatest challenges for the CHI community.

To address these challenges, Fisher [10] and her research group are co-designing social media and mobile tools with Syrian women to support and empower them through several initiatives in Jordan. In this work, starting from the results of ongoing fieldwork carried out in Jordan in the last months, we identify some research challenges in the field and present a preliminary proof of concept of a messaging system which addresses one of the identified challenges, creating a bridge between human and computer science.

3 MOTIVATION ANALYSIS

Ongoing fieldwork conducted by Fisher and colleagues focuses on communication and privacy (hurma), which are ruled by 1) secular/government law; 2) Islamic/Sharia law that includes relations, information sharing, speech and dress; and 3) tribal law/cultural norms prizing collective-family honor (Sharaf) over the individual, passed generationally via blood lines and across borders [1, 2]. At any time all are at play, interpreted locally and regionally. With whom can a woman be friends, chat privately, or follow? What can she post or use in her profile? How can she advertise a business or seek help? Concepts of halal (permissible) and hijab (curtain) dictate friending immediate family and not sharing photos of herself; often a male guardian's permission for an account, sometimes providing him the password. Some accounts are in a brother or son's name with photo, which can be confusing for westerners viewing posts and messaging. "Haram" is illicit behavior; LGBTQ is especially haram by secular and tribal law, needing great privacy protection. Other vulnerabilities are DBV, human trafficking, early marriage, muta'a (temporary pleasure marriage), divorce, and enterprise. For Syrians, war adds heightened privacy needs to protect identity and location—accounts in multiple identities, strict friendship and posting rules, and strategies for vetting sources and information. Using Fisher's Humanitarian Research "People First, Data Second" [10], fieldwork comprising codesign workshops, group discussions, and participant observation using scenarios and Quran cards - based on prior fieldwork and in collaboration with Imams - is yielding deep understanding of women' social media experiences, their digital literacy needs, and feature recommendations based on Arab-Islamic culture. Specifically, the fieldwork with women and men, both Syrian and Jordanian and predominantly aged 20-30, across villages and urban centers in Jordan is addressing:

(1) digital literacy: what DL means, how females use social media, their goals, and online experiences;
(2) females' online practices to protect privacy and against online perils of mis-, dis- and malinformation;
(3) how women self-define technological needs and practices;
(4) the roles and impacts of secular, Islamic and tribal law, and social/cultural factors on social media use;
(5) how female users assess social media features and content for privacy protection and capability;
(6) how female users would design and refine social media and digital literacy support to ensure safety and privacy preservation and promote capability.

While the fieldwork methodology and in-depth findings are discussed in other reports (in submission), analysis of data collected to-date with over 100 women and 25 men indicates the following: 1) it is extremely important to women and men that their online behavior complies with tenets of Islam and Arab culture, which are currently largely unsupported by mainstream social media; 2) women and men have experienced much dis and mis-information, hacking, trolling, and blackmail that have negatively affected their lives; and 3) that social media safety can be strengthened by integrating specific Surah and Aya from the Quran as well as hadith and Arab culture to base feature design. Specific recommendations with examples from the fieldwork (using aliases) include:

- Warning for Sensitive Content (F1)
  34 year-old Amira said she and her friends dislike the sight of blood and seeing people hurt or trauma, especially as these photos often trigger painful memories. Sometimes such photos can be of a family or location they know, which makes such images even more painful. She and her friends also do not want to see photos of war or hurt animals.
- Being informed if you are going to send Explicit Content (F2)
  A young woman often exchanges private photos with her sisters in Syria and Turkey of them wearing short-sleeve summer dress and without hijab. She worries someone will hack their phone and photos and would like to be reminded before sending a sensitive photo.
- Malicious URLs Detection (F3)
  24 year-old Nour clicked on a facebook link about financial support for education; her account was immediately hacked. Other young women shared stories of clicking on links about immigration that led to or included pornography.
- Forwarding Control of Personal Content (F4)
  A friend of 21 year-old Maria posted a private photo of Maria on Facebook. It caused a fight between Maria and her friend. They resolved the problem by hiding the photo.
- Detection of Blackmails (F5)
  Blackmail is often preceded by hacking as 32 year-old Haya explained and it has happened to herself and many friends, both female and male. First your Facebook messenger account is hacked and the person messages you (using your profile and photo), saying to sending them money via a Whatsapp number to make them leave your account and not delete or share your photos with others, or message your friends, etc.
  Blackmail also occurs when a person messages a woman saying he will inform her father or family that an incident occurred unless she agrees to something he wants. Leana explained a man messaged her saying he would tell her father that she invited him to her apartment unless she agreed to actually see him.
- Blur Children's Faces (F6)
  Before Eid Adha (the Islam celebration before Haj), a woman posted a photo of her two small daughters on Facebook. According to Islam and Arab culture, when a person sees something positive, it is expected for the person to say

Table 1. Evaluation of the privacy-preserving Features

Score	Feature
4.5	End-to-End Encryption (F9)
4.4	Malicious URLs Detection (F3)
4.3	Forwarding Control of Personal Content (F4)
4.1	Detection of Blackmails (F5)
4.0	Detection of Deepfakes (F8)
3.9	Identifying Personal Photos posted by someone else (F7)
3.8	Warning for Sensitive Content (F1)
3.2	Being informed if you are going to send Explicit Content (F2)
3.1	Blur Children's Faces (F6)

Mashallah ("blessed by God"). Few of her Facebook friends, in the woman's mind, said "Mashallah" and thus one of her daughters became sick. She then deleted the photo, and in future said she will hide her daughters' faces.

- Identifying Personal Photos posted by someone else (F7)
  A young woman who works at an NGO was in private and small groups photo and a person posted the images without asking for consent.
  A 42 year-old Syrian woman has grown a successful home business and did a few interviews with European media who promised not to show her face—only from the back of her head. Instead, they posted a video showing her face. Shortly after, her family in Syria received whatsapp and Telegram messages from strangers asking if they know her and her location.
  Young woman was at beauty salon with her friends. Someone they do no know posted photos of them without hijab on a whataspp group.
- Detection of Deepfakes (F8)
  Photos of a male friend were photoshopped and posted on a pornography site. He did not learn about it for several months. It caused him and his family much problems.
- End-to-End Encryption (F9)
  Whatsapp was explained as the preferred messaging app because of end-to-end encryption, which everyone highly values. The difficulty is people use other platforms such as Facebook and Instagram very frequently and forget they do not have the same protections.

Although these features have been identified interacting with Syrian and Jordania women (and men), they are indeed important for every user. To prove this claim, we have administered a questionnaire to 102 Italian computer science students. In particular, participants were asked to evaluate from 1 (*very little*) to 5 (*very much*) the usefulness of each of the aforementioned features. The results are reported in Table 1, ranked in descendent order of score.

4 ANALYSIS OF THE FEATURES

In this section, we analyze the privacy-preserving features identified in Section 3 and present some technological approaches, except for end-to-end encryption (F9) which can be implemented by any messaging platform.

4.1 Warning for Sensitive Content (F1 and F2)

Seeing disturbing content while scrolling the feed of our favourite social network or receiving unwanted explicit content on a messaging application is quite frequent. However, uncensored explicit content may affect people's well-being, especially those with past traumatic experiences or medical conditions such as Post-Traumatic Stress Disorder (PTSD). In our case study, Syrian refugees in Jordan have experienced trauma due to the atrocities of war [10]. Asking for users' explicit consent before showing potentially disturbing content allows them not to be exposed to additional trauma [29].

One possible approach is the employment of trigger warnings, which hide the content until users do a predefined action, such as a click or other gestures. Yet, identifying explicit and sensitive content may be tricky as people have different perceptiveness of sensitivity to the definition of disturbing content. For instance, in the Arab culture, if content is spread outside the users' intended audience, it can bring shame and embarrassment to users and their family [2]. Furthermore, defining a policy which can strike a balance between free expression and a safe platform is even more complicated [16]. For instance, consider contents such as drugs, blood, etc.; they could be allowed in case of community discussion about the effects of drugs or informative video about surgery, etc.

Several APIs for content moderation are available out of the box for developers on the numerous cloud services. Some representative examples are the Google Safe Search API and Amazon Rekognition. They differ in the type of violations they can recognize and the information they provide to the developer. For instance, Google Safe Search API provide 5 different levels of likelihood (from VERY_UNLIKELY to VERY_LIKELY) for 5 categories of content (adult, spoof, medical, violence, racy). Instead, Amazon Rekognition can detect 10 top-level categories and more than 30 subcategories of inappropriate or offensive content, providing a confidence value for each label. Such a value may be exploited to discriminate between different levels of explicit content, yet testing is needed. Since these solutions run in the cloud and the delivery time is a crucial factor in messaging platform, an evaluation of the response time is needed.

These solutions are not available on the device, raising some privacy concerns since contents need to be sent in clear text through the network to the external service. In the case of an end-to-end encrypted platform, adopting one of these solutions is not possible since the content would be encrypted between sender and receiver. Still, an ML-based custom solution deployed on the device is compatible with end-to-end encryption and does not suffer from any issues related to network latency. Yet, a proper amount of data is needed to train the system. Furthermore, also energy consumption needs to be considered since it is a crucial factor for mobile devices [8].

The technological approaches described in this section can also be employed to inform users when they are going to send explicit content. Similar approaches hold even when considering media types other than images. E.g., the content moderation API of Amazon Rekognition works even with video. Furthermore, it is possible to use services (or custom solutions, also on the device) to analyse texts or even audio, so tasks such as hate speech detection become feasible.

4.2 Malicious URLs Detection (F3)

Web pages can contain malicious content (e.g., malware, phishing, etc.) for users, jeopardizing their experience and causing social and economic issues. For instance, consider that criminals can employ phishing attacks to lure people into transfer money or install malware, which can then steal users' personal data. To visit these web pages, users have to click on an URL, implicitly evaluating the associated risks. Yet, even heavy technology users sometimes encounter difficulties in accurately predicting the destination, and consequently the safety, of an URL [5]. In our scenario, letting people alone evaluate the safety of URLs would be unfeasible since their low digital literacy [10]. Hence, without denying that education is essential, an automatic solution that supports users in making an informed decision is necessary.

Furthermore, some web links, even without pointing to malicious content, can lead to disturbing and/or sensitive material, such as adult content or violence. This deserves a solution for the same reasons discussed in Section 4.1.

Solutions for the automatic detection of malicious URLs have been proposed in the literature. For instance, Abutaha *et al.* [3] proposed a technique to detect phishing based on lexical analysis and ML. Instead, Rajalakshmi *et al.* [26] presented a deep classifier for detecting URLs pointing to adult content. Ma *et al.* [20] proposed an online learning algorithm for detecting malicious URLs, acknowledging that the Web is dynamic and an algorithm that rapidly adapts to its changes is necessary. Althobaiti *et al.* [6] investigated the usage of a report that exploits existing information about an URL (e.g., domain age, domain popularity, presence of manipulation tricks, etc.) to support users while judging its safety. In this way, users can access information usually available only to experts. In addition, the report can also help educating users and improving their digital literacy [6].

4.3 Forwarding Control of Personal Content (F4)

The spreading of sensitive content without the owner's consent can have serious consequences. This is even more true in our context. For instance, consider that if a female's photo is misused, it will bring embarrassment to her family [2]. Another noteworthy case is sexting, where self-generated nude images can be spread by the receiver (e.g., revenge porn). Furthermore, the commonly used messaging applications allow users to send any content, even private ones, to anyone, thus exposing them to the possible consequences [12]. For these reasons, Franco *et al.* [12] have proposed some guidelines for developers to build safer messaging applications than the current ones for sexting, thanks to a forwarding control feature for images. Their approach can be valuable even in cases different from sexting.

4.4 Detection of Blackmails (F5)

Blackmailing can be defined as the threat of sharing damaging private information, which is presumed to be injurious to the victim [19, 24]. One representative example is sextortion, defined as "the threatened dissemination of explicit, intimate, or embarrassing images of a sexual nature without consent, usually for the purpose of procuring additional images, sexual acts, money or something else" [24]. Patchin *et al.* [24] estimated that about 5% of US middle and high school students experienced sextortion. This finding is in line with the one of Kopecký, who demonstrated that 6-8% of teenagers in the Czech Republic reported serious cases of blackmail [19]. The consequences of sextortion and blackmailing are very dangerous, as demonstrated by the numerous cases of suicide, just to mention one example.

Yet, despite the dangerousness of this online activity, little is known about possible technological approaches to prevent it, recognize it or support victims. Detecting blackmails is a challenging technological problem since they are often carried out in several stages [19], thus potential solutions need to consider different aspects simultaneously. Still, sentimental analysis techniques may be exploited to detect texts containing threats. In this way, once a message is classified as blackmail, the platforms can show an alert to inform users, suggesting a safe way to react, which usually consists of not answering at all. Other solutions can help in preventing the extortion itself. For instance, changing a part of the user interface when the camera is on may increase the user's awareness.

4.5 Blur Children's Faces (F6)

Posting photos depicting someone else on social media is not legal in many countries. Even more dangerous is parental sharing of images of their children. Parents rarely ask for children's consent, even because children are too young to provide permission. Yet, such photos may cause children's embarrassment in the future, be no longer in the parents' control, or be misused by other social media users [7].

One possible technological approach combines face detection and age estimation, exploiting the body of knowledge about deep learning techniques for solving these tasks. Then, without denying that educating users about the privacy drawbacks of sharing such content is essential, faces with an estimated age below a fixed threshold (e.g., 18) could be blurred. Yet, other behaviours can be implemented; for instance, culturally-tailored comments or reactions can be proposed to users. Indeed, culture may influence how users have to react within a social media. For instance, for Muslim users a reaction with "*Mashallah*" as answer to something positive is preferred to the standard "*like*", as mentioned in Section 3. This means that the common types of reaction, i.e, *like, dislike, love, smile*, etc., do not work well for everyone and that the definition of a set of reactions that considers also Islam and Arab culture embodies an open issue.

4.6 Identifying Personal Photos posted by someone else (F7)

On social media, anyone can publish a photo depicting someone else without his/her consent. The possibility to block this action can improve the platform's safety and privacy. To do so, we first need to collect at least one face image for each user. Then, once someone sends an image, the system can compare it with the known images/faces. Such a comparison can be done through computer vision cloud services (e.g., Amazon Rekognition) or a custom solution [28]. The platform can store either images or embeddings. However, privacy concerns need to be considered during the design phase, also for the compliance with current regulations (e.g., GDPR, DMCA). In addition, verifying the identity of people (how can we be sure that an image loaded by a user depicts himself/herself?) may be problematic, or even not possible from an ethical standpoint. Furthermore computation time and impact on user experience need to be evaluated.

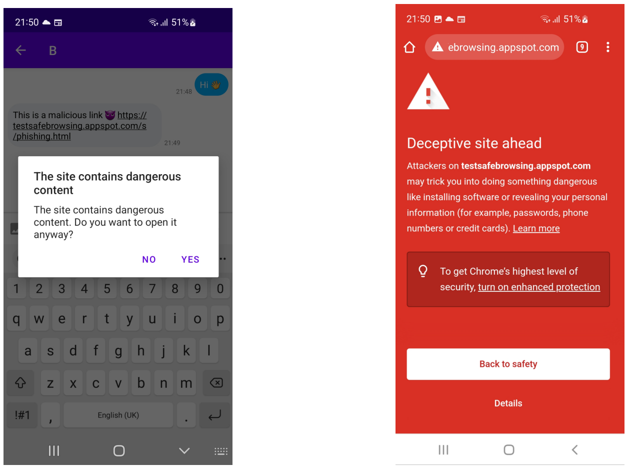
4.7 Detection of Deepfakes (F8)

Deepfakes are artificial media generated by deep neural networks where one's personal voice, image or video is replaced by the one of someone else, thus appearing real, created by leveraging on AI and ML techniques [22]. Although some beneficial applications in the fashion and movie industry, they can have numerous harmful usages with important social implications [13, 15, 22]. For instance, malicious users could impersonate an identity, spread disinformation, swap a victim's face onto the body of a porn actor, blackmail someone, or even build dynamic online profiles. Consequently, these attacks cause political, economic, psychological, and even physical harm. Furthermore, if media can be easily manipulated, it becomes difficult to verify their authenticity, thus reducing society's trust in visual media [15]. In our context, deepfakes can be even more harmful because of the trauma and the low digital literacy [10].

Several works have proposed solutions for detecting deepfakes. Some of them exploit artefacts generated during the creation process of deepfakes, whereas others are based on the analysis of physiological cues of human faces [16]. Mirsky *et al.* presented a comprehensive review of techniques for detecting deepfakes in [22]. Yet, it is well known that ML-based can make wrong predictions. Therefore, some works propose techniques for preventing the creation of deepfakes. For instance, data provenance can be tracked through blockchain-based solutions, and perturbation can be added to images, corrupting deepfakes networks, which will no more be able to locate faces properly.

5 RESEARCH CHALLENGES: PRELIMINARY RESULTS

To address the identified research challenges and as a first step toward the goal of building safer and more trustable social media, we have focused on the most important features as ranked in Section 3. As the end-to-end encryption is well known and several social media already implement it, we have considered the second feature in order of importance in Table 1, the detection of malicious URLs, and implemented it as a proof of concept in a custom messaging system.



Hi 👋

This is a malicious link 😈 https://testsafebrowsing.appspot.com/s/phishing.html

The site contains dangerous content

The site contains dangerous content. Do you want to open it anyway?

Deceptive site ahead

Attackers on **testsafebrowsing.appspot.com** may trick you into doing something dangerous like installing software or revealing your personal information (for example, passwords, phone numbers or credit cards). Learn more

To get Chrome’s highest level of security, turn on enhanced protection

(a) Request for Explicit Confirmation before Opening an Unsafe URL

(b) Warning Message of Browser for a Malicious URL

Fig. 1. Proof-of-Concept of the Feature for Detecting Malicious URLs

Our platform is composed of an Android application, whereas its backend is based on Google Firebase. In particular, Firebase Storage store images, users’ data and messages are saved in Firebase Database (Cloud Firestore), while Firebase Functions contains the logic which is coded in TypeScript, using the NodeJS framework.

Our application can detect malicious URLs. When a user clicks on an URL, the system checks whether it is safe or not. If no threats are detected, the system opens the link as usual. Otherwise, the system shows an alert that informs the user and requires an explicit confirmation before opening the URL, as shown in Figure 1a.

The platform employs the Google SafetyNet Safe Browsing API for checking the safeness of an URL, which internally implements a client for Google Safe Browsing Network Protocol v4. Basically, it checks whether an URL has been marked as a known threat by Google. In other words, each web link is compared with a list of malicious URLs.

Such an implementation performs a network request to verify the safeness of the URL, yet the response is cached for a fixed period. Although it is simple, privacy concerns and response time can be drawbacks. Still, it is possible to directly employ Google Safe Browsing API and perform the computation locally on the device. Indeed, our application could periodically download an encrypted version of Safe Browsing lists and check the safeness of an URL client-side. Clearly, by employing this approach, the server never knows the URLs queried by the clients and the response time is almost zero. Yet, the implementation is more complex. Google provides an example but it was not suitable for our purposes. Therefore, we decided to implement the version that performs a network request, as it is implemented directly in the Android OS, allowing us to build our prototype quickly. A future improvement will be implementing a client for the latter approach since it would be ideal in a high network delay scenario. On the other hand, opening a website would be difficult in such a scenario, so also the possible damage of a malicious website is reduced.

In Figure 1b, we note that when the user decides to open a malicious link despite our alert, even the browser displays a warning. This happens because both the considered browser and our system employ the same API for detecting malicious links. Yet, our feature can be useful as a first barrier and when users employ web browsers not performing such a verification. Furthermore, if an URL is detected as unsafe by such an API, we will be sure about the correctness of this detection. On the other hand, if a link is classified as safe, we will not know whether this is true, since our solution only compare an URL with a list of known malicious ones. Employing more complex approaches, such as the ones described in Section 4.2, would improve the effectiveness of this feature.

6 CONCLUSION

People, especially vulnerable ones such as some Arabic women, face challenges to their privacy in the everyday use of social media, which can potentially have harmful consequences, such as job loss, social embarrassment, imprisonment, etc. Yet, these platforms are a lifeline for Arabic females, allowing them to communicate with their families, see the lives of their friends, or read the news, just to mention some examples.

In this paper, starting from the findings emerging from an ongoing fieldwork in Jordan, we identified features for improving safety on social media platforms that may be crucial to Arab Muslim populations, but also generally useful to anyone. We have described each of these feature and the issue they address. We have also discussed the possible technological approaches to implement them. As a first step toward safer and more trustable messaging platforms, we have developed a proof-of-concept of a messaging system able to detect malicious URLs received by users.

Yet, this is only a first step toward the goal of protecting Arabic-Muslim women through safer social media. Therefore, we are extending our research in several directions. First of all, more features may be identified and analyzed through other further fieldwork, which includes males, in Jordan. A comparison between the data collected in Jordan and data gathered in Western countries (e.g., Italy) would be particularly interesting. Furthermore, we plan to extend our proof-of-concept by implementing some of the other identified privacy-preserving features. Then we would like to perform a test campaign, evaluating both the efficacy and the usability of our platform/features. Forthcoming papers also focus specifically on analyzing the fieldwork data with regard to the Quran and Arab cultural norms.

ACKNOWLEDGMENTS

This work has been partially funded by the Department of Mathematics of the University of Padua (BIRD191227 project). Other supports include Google Research Inclusion Award 2020 and Facebook Integrity Award 2021.

REFERENCES

[1] Norah Abokhodair, Adam Hodges, and Sarah Vieweg. 2017. Photo Sharing in the Arab Gulf: Expressing the Collective and Autonomous Selves. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 696–711. https://doi.org/10.1145/2998181.2998338

[2] Norah Abokhodair and Sarah Vieweg. 2016. Privacy & Social Media in the Context of the Arab Gulf. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 672–683. https://doi.org/10.1145/2901790.2901873

[3] Mohammed Abutaha, Mohammad Ababneh, Khaled Mahmoud, and Sherenaz Al-Haj Baddar. 2021. URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis. In *2021 12th International Conference on Information and Communication Systems (ICICS)*. 147–152. https://doi.org/10.1109/ICICS52457.2021.9464539

[4] Aya Akkawi and Ayat Nashwan. 2019. My Name is Salma and I am a Victim of Honor Crimes: (Re) conceptualizing Honor Killing and Stigma against Women. (01 2019).

[5] Sara Albakry, Kami Vaniea, and Maria K. Wolters. 2020. *What is This URL's Destination? Empirical Evaluation of Users' URL Reading.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376168

[6] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. 2021. I Don't Need an Expert! Making URL Phishing Features Human Comprehensible. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 695, 17 pages. https://doi.org/10.1145/3411764.3445574

[7] Mary Jean Amon, Nika Kartvelishvili, Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia. 2022. Sharenting and Children's Privacy in the United States: Parenting Style, Practices, and Perspectives on Sharing Young Children's Photos on Social Media. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 116 (apr 2022), 30 pages. https://doi.org/10.1145/3512963

[8] Matteo Ciman and Ombretta Gaggi. 2017. An empirical analysis of energy consumption of cross-platform frameworks for mobile development. *Pervasive and Mobile Computing* 39 (2017), 214–230. https://doi.org/10.1016/j.pmcj.2016.10.004

[9] Karen Fisher. 2018. *Digital Lifeline?: ICTs for Refugees and Displaced Persons.* Carleen Maitland and Sandra Braman Ed., The MIT Press, Chapter Information worlds of refugees, 79–112.

[10] Karen Fisher. 2022. People First, Data Second: A Humanitarian Research Framework for Fieldwork with Refugees by War Zones. *Computer Supported Cooperative Work (CSCW)* (03 2022), 1–61. https://doi.org/10.1007/s10606-022-09425-8

[11] United Nations High Commissioner for Refugees (UNHCR). 2021. Midyear Trends 2021. Geneva, Switzerland: Office of the United Nations High Commissioner for Refugees. https://www.unhcr.org/statistics/unhcrstats/618ae4694/mid-year-trends-2021.html

[12] Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. 2022. Improving Sexting Safety through Media Forwarding Control. In *2022 IEEE 19th Annual Consumer Communications Networking Conference (CCNC).* 1–6. https://doi.org/10.1109/CCNC49033.2022.9700555

[13] Dilrukshi Gamage, Piyush Ghasiya, Vamshi Bonagiri, Mark E. Whiting, and Kazutoshi Sasahara. 2022. Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22).* Association for Computing Machinery, New York, NY, USA, Article 103, 19 pages. https://doi.org/10.1145/3491102.3517446

[14] Arup Kumar Ghosh, Karla Badillo-Urquiola, Mary Beth Rosson, Heng Xu, John M. Carroll, and Pamela J. Wisniewski. 2018. A Matter of Control or Safety? Examining Parental Use of Technical Monitoring Apps on Teens' Mobile Devices. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173768

[15] Samuel Greengard. 2019. Will Deepfakes Do Deep Damage? *Commun. ACM* 63, 1 (dec 2019), 17–19. https://doi.org/10.1145/3371409

[16] Alon Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2022. Preserving Integrity in Online Social Networks. *Commun. ACM* 65, 2 (jan 2022), 92–98. https://doi.org/10.1145/3462671

[17] Husseini. 2020. FB live abuse plea gains ground. jordantimes.com/news/local/womans-facebook-live-abuse-plea-gains-ground

[18] Simon Kemp. 2022. Digital Jordan: 2022. https://datareportal.com/reports/digital-2022-jordan

[19] Kamil Kopecký. 2017. Online blackmail of Czech children focused on so-called "sextortion" (analysis of culprit and victim behaviors). *Telematics and Informatics* 34, 1 (2017), 11–19. https://doi.org/10.1016/j.tele.2016.04.004

[20] Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. 2011. Learning to Detect Malicious URLs. *ACM Trans. Intell. Syst. Technol.* 2, 3, Article 30 (may 2011), 24 pages. https://doi.org/10.1145/1961189.1961202

[21] Brenna McNally, Priya Kumar, Chelsea Hordatt, Matthew Louis Mauriello, Shalmali Naik, Leyla Norooz, Alazandra Shorter, Evan Golub, and Allison Druin. 2018. Co-Designing Mobile Online Safety Applications with Children. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18).* Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3173574.3174097

[22] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (jan 2021), 41 pages. https://doi.org/10.1145/3425780

[23] Samia Oukemeni, Helena Rifà-Pous, and Joan Manuel Marquès Puig. 2019. Privacy Analysis on Microblogging Online Social Networks: A Survey. *ACM Comput. Surv.* 52, 3, Article 60 (jun 2019), 36 pages. https://doi.org/10.1145/3321481

[24] Justin W. Patchin and Sameer Hinduja. 2020. Sextortion Among Adolescents: Results From a National Survey of U.S. Youth. *Sexual Abuse* 32, 1 (2020), 30–54. https://doi.org/10.1177/1079063218800469 arXiv:https://doi.org/10.1177/1079063218800469 PMID: 30264657.

[25] Damian Radcliffe. 2021. SM trends in MENA. https://wan-ifra.org/2021/06/tiktok-trumps-snapchat-social-media-trends-in-mena-in-2020

[26] R. Rajalakshmi, Joel Raymann, Aneesh Prabu, and Chandrabose Aravindan. 2019. Deep URL: Design of Adult URL Classifier Using Deep Neural Network. In *Proceedings of the International Conference on Advanced Information Science and System* (Singapore, Singapore) *(AISS '19).* Association for Computing Machinery, New York, NY, USA, Article 20, 5 pages. https://doi.org/10.1145/3373477.3373497

[27] Susan Razzaz. 2017. *A Challenging Market becomes More Challenging: Jordanian Workers, Migrant Workers and Refugees in the Jordanian Labour Market.* International Labour Organization. www.ilo.org/wcmsp5/groups/public/---arabstates/---ro-beirut/documents/publication/wcms_556931.pdf

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 815–823. https://doi.org/10.1109/CVPR.2015.7298682

[29] Manuka Stratta, Julia Park, and Cooper deNicola. 2020. Automated Content Warnings for Sensitive Posts. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20).* Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383029

[30] Zeynep Şahin Mencütek and Ayat J. J. Nashwan. 2021. Employment of Syrian refugees in Jordan: challenges and opportunities. *Journal of Ethnic & Cultural Diversity in Social Work* 30, 6 (2021), 500–522. https://doi.org/10.1080/15313204.2020.1753614 arXiv:https://doi.org/10.1080/15313204.2020.1753614