DEMOCRACY REPORTING INTERNATIONAL

APRIL 2022 • OCTOBER 2022

ONLINE DISINFORMATION AND HATE SPEECH IN THE MENA REGION

REGIONAL TRENDS AND LOCAL NARRATIVES

Warning:

Social media monitoring reports contain potentially disturbing content that may be distressing for some readers.

Democracy Reporting International is sharing this content only for scientific and research purposes.

This second report has been produced by DRI and its partners for the project "Words Matter". The report covers the period from April 2022 to October 2022:

الكلمة تفرق
Words Matter

DRI Partners

RASED راصد
مركز الحياة AL HAYAT CENTER
مراقبون Mourakiboun
ممارات
SUDIA
Jordan Open Source Association

Supported by

german cooperation
DEUTSCHE ZUSAMMENARBEIT

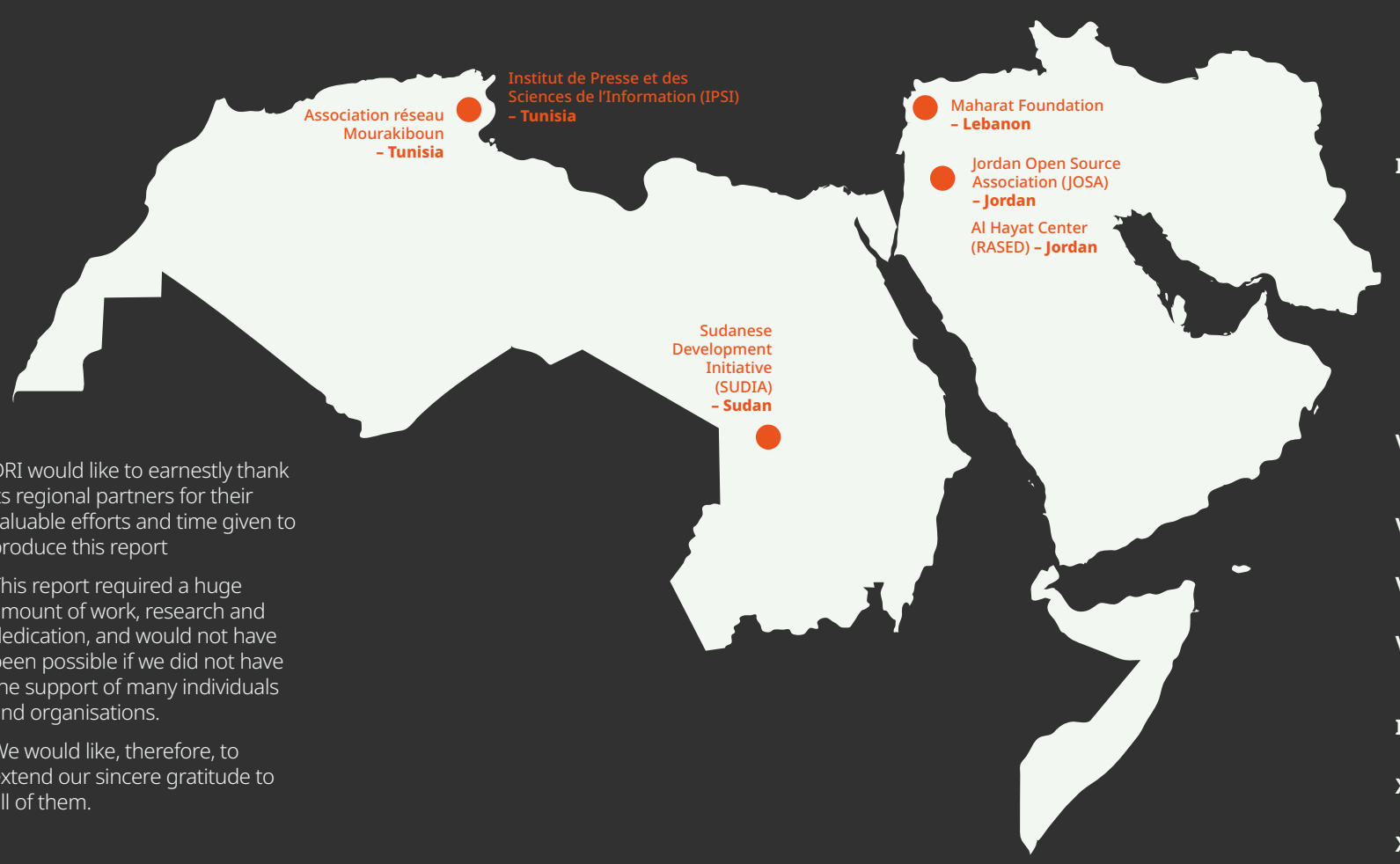
February 2023

Acknowledgments

DRI would like to earnestly thank its regional partners for their valuable efforts and time given to produce this report

This report required a huge amount of work, research and dedication, and would not have been possible if we did not have the support of many individuals and organisations.

We would like, therefore, to extend our sincere gratitude to all of them.



Index

Executive Summary

This report is the second of four regional social media monitoring reports to be produced as part of the "Words Matter" project, focusing on countering disinformation and hate speech in the MENA region. The project aims to strengthen the safeguarding of democratic processes and the resilience of societies in the region to online disinformation and hate speech. It builds on the assumption that civil society actors, including the media, are essential to monitoring, understanding, and raising awareness of what debates and discourses are occurring online.

This report investigates online disinformation and hate speech trends during key national democratic processes in three countries in the MENA region (Tunisia, Jordan and Lebanon), including findings from

- a constitutional referendum in Tunisia;
- parliamentary elections in Lebanon;
- and the online debate that took place in Jordan around the publication of a Childrens' Rights bill (draft law).

It also develops national and regional recommendations intended for civil society organisations (CSOs), researchers and social media platforms to improve moderation and promote transparent regulations to control online disinformation and hate speech.

Project partners used several tools to monitor discourse on social media, including Meta's CrowdTangle [1], Twitter API [2], TweetDeck from Twitter, and DRI's own Digital Democracy Monitor Toolkit in Arabic.

First, Tunisia's "Lab'TRACK" project presents its findings during the online campaign preceding the national referendum on a new constitution in the country, held on 25 July 2022. This project joins the efforts of Mourakiboun and the Institute of Press and Information Sciences (IPSI – Institut de Presse et des Sciences de l'Information). It provides insights into the growing use of memes, humour and Facebook live video streaming to spread disinformation in a manner that can escape detection and relies increasingly on organic dissemination.

Second, the Maharat Foundation presents its monitoring conducted in Lebanon during the campaign for the parliamentary elections of May 2022 and the period immediately after, which shows the high level of negative content in online campaigns, the extensive use of fake Twitter accounts, and a significant level of hate speech against women politicians.

Third, Al Hayat Rased, from Jordan, presents the research it conducted on online abuse and harassment of supporters of draft legislation on children rights, with a special focus on gender-based hate speech, which draws heavily on religious references and allusions to family origins to denigrate women.

Under the monitoring period covered by this report, we have detected several regional trends and tactics that were used by different actors to spread hate speech and disinformation to influence the narrative in the digital public spheres. The first regional trend that emerged during our monitoring was online gender-based violence, which is prevalent on social media platforms, especially during times of elections and crises. Our findings focus on three countries: Lebanon, Jordan, and Sudan. In Lebanon, research found that 43% of the social media accounts of 100 active women candidates showed various forms of online violence against women. This was mostly in the form of psychological violence, with the remainder shared equally between sexual violence and comments on their age and appearance. In Jordan, online abuse against those who supported the draft law on Children's Rights was found to be extremely gendered, with men being attacked for their social status and women being attacked on many levels including their foreign origins, rights, and place in political life. In Sudan, misogyny, and gender-based violence is a pattern on social media platforms, as well as on the ground during political demonstrations.

The second regional trend is religious-based hate speech. We explore this trend in two countries: Lebanon and Jordan. In Lebanon, Maharat Foundation documented the use of shallow or cheap fakes, which are audio-video manipulations to spread hateful content based on religious grounds. They found that Lebanese MPs who adopted a civil marriage law have been subjected to a religious hate campaign on social media, led by Sunni Muslim Sheikhs. In Jordan, Al-Hayat-Rased Organization documented the use of negative religious discourse against women at all levels, regardless of whether they were wearing the Hijab or not. The attacks were carried out against those

[1] CrowdTangle: https://dataforgood.facebook.com/dfg/tools/crowd-tangle

[2] Twitter API: https://developer.twitter.com/en/docs/twitter-api

who publicly supported the Children's Rights Bill on social media. This was manifested in the criticism of former MP Rola Al-Hroub for not wearing the hijab and for her comments about law and Sharia. This type of hate speech was used to manipulate and spread hateful messages on social media.

The third regional trend is the use of emotionally induced discourse in political manipulation. In Tunisia, coordinated disinformation content on social media that bears a strong emotional charge is used to amplify its dissemination and manipulate public sentiment. Anger is used as a tool for making misinformation go viral. In Lebanon, during parliamentary elections, 49.5% of online discourse was found to be based on the manipulation of sentiments. Traditional political forces rely more on emotional rhetoric to strengthen partisan and sectarian affiliation and evoke conspiracy theories to demonize opponents.

The fourth trend is coordinated disinformation campaigns during major national political events. In Tunisia, political actors use Facebook's live function to broadcast simultaneous streams on several pages to reach as many followers as possible and spread hate speech and misleading information. This pattern is also very effective when it comes to establishing direct interaction with specific communities. These coordinated campaigns spread misinformation and manipulate public opinion during critical political events.

The monitoring also uncovered various tactics that are used by many actors to spread hate speech and online manipulation in the region. One of the main tactics is the use of satire through political memes. In Tunisia, we have documented the use of memes to spread harmful content, with a regular target being politician Abir Moussi who opposed the referendum on the constitution that was initiated by President Kais Saied. Another tactic that is used is the coordinated live videos which notifies the page followers in advance which generates higher engagement than pre-recorded videos. These live videos are often broadcasted simultaneously on multiple pages making it harder to detect and moderate with automatic content moderation filters. The third tactic was creating posts that would ignite a lot of hateful comments as we have found in many cases that comments on social media platforms contain more hateful content than the posts as it was observed by our partners in Lebanon and Jordan. The last tactic that we have uncovered was the activation of fake social media accounts to amplify specific political messages and participate in defamation campaigns against politicians. These accounts are only activated during periods of elections and democratic transitions.

The report includes country-specific and regional recommendations, based on the social media monitoring efforts and observations from our partners.

Recommendations at the regional level:

For political parties and movements

- The development of an internal code of conduct to prevent members from engaging in hate speech and disinformation on social media.

For traditional media:

- Engagement with civil society and peer organisations in the region to develop and commit to reporting standards and ethics that do not promote hate speech.

For the community of researchers in the MENA region,

- Improved collaboration between computational researchers, digital rights activists, sociologists and other analysts, on a regional scale.
- Continued vigilance in relation to new forms of disinformation and to the use of AI in generating and spreading disinformation and hate speech through synthetic media and deepfakes.

For Tech Platforms

- Greater investment in content moderation, by: (i) hiring more local staff from different parts of the region to perform human content moderation; (ii) investing in language modules of algorithms to detect hate speech in Arabic and local dialects in the region; and (iii) enhancing their reporting mechanisms.
- Collaboration with researchers in the MENA region to develop hate speech lexicons in different dialects to detect different forms of harmful content in the region.
- Working closely with civil society and electoral bodies to introduce and explain their community guidelines and content moderation policies.
- Vigilance in relation to new tactics of bad actors, for example, the spread of hate speech in the comments section on Facebook, the use of text overlay in images (memes), and leaks and doxing the private information of political candidates. Further develop technical capabilities to monitor harmful content shared via live videos.
- Support to independent media platforms promoting safe spaces for women active in the political field, and digital media literacy initiatives in general.
- Lifting restrictions on new users of monitoring application tools, such as CrowdTangle, and facilitating access to information for CSOs and research and academic institutions.[3]

Civil Society

- Strengthening regional and regular collaboration and engagement with governmental bodies, legislators, specialised independent authorities, and tech companies.

[3] Since January 2022, Meta has stopped the right of new users to use the CrowdTangle tool to search social media.

- The development of countering strategies and regional coalitions to monitor, document, analyse ,and respond to harmful and dangerous content on social networking sites.
- The launch of more initiatives for information verification, networking, and exchange of experiences.
- The organisation of national workshops on legislative amendments or improved and transparent enforcement, to better deal with the challenges of hate speech, disinformation, and false news.
- The formation of ad hoc multi-disciplinary bodies to better monitor and demand corporate accountability from tech companies.
- The creation of helplines to support victims of online gender-based violence and assist them in reporting harmful content to social media companies.
- The promotion of digital literacy among social media users.

Words Matter partners faced some challenges in the monitoring of online abuse:

1. Most of the tools used in collecting data, such as Twitter API and CrowdTangle, are not suitable for social media monitoring in Arabic, because of the variations between dialects and standard Arabic. They are even less suited for other local languages spoken in the MENA region.
2. Our researchers documented difficulties choosing specific locations on CrowdTangle.
3. The news about Meta potentially planning to stop supporting CrowdTangle hampers the work of social media researchers and fact-checkers who are building their methodology to research and collect data from that resource.
4. There were some technical issues with ExportComment tools, meaning that multiple requests had to be sent to the tool, which delayed data collection.
5. Different national contexts and different issues surveyed prevented Words Matter network partners from using a unified methodology to monitor hate speech and disinformation. This made support to partners more time-consuming and cross learning less fruitful.

The report also contains an interview with the social media researcher Lena-Maria Böswald, discussing the emerging threats to online discourse and the shaping of public opinion by the use of artificial intelligence to generate and spread disinformation.

Broader, contextualised research continues to be needed to explore the environment, whether legal, technical, or societal, that enables hate speech, in general, and gender-based violence on Arabic-speaking social media platforms (including beyond Facebook and Twitter), in particular. This will be a priority of the next report.

Introduction

DRI works with partner organisations from four countries (Jordan, Lebanon, Sudan and Tunisia), strengthening local capacities to monitor and analyse online disinformation and hate speech, while building a regional network to allow for comparative analysis and peer learning.

This report is the second of a series of four produced by the network of partners. It aims to:

- Analyse disinformation and hate speech during key national democratic processes and political events, to shed light on behaviours, patterns and streams, with a particular focus on organic dissemination and on gender-based harassment and violence; and
- Propose national and regional measures to counteract online disinformation and hate speech

With different levels of their progress in each country, the report presents the work accomplished by DRI's partners as follows:

First, the "Lab'TRACK" project presents its work conducted during the online campaign preceding the national referendum on a new constitution in Tunisia, held on 25 July 2022. This project joins the efforts of Mourakiboun and the Institute of Press and Information Sciences (IPSI - Institut de Presse et des Sciences de l'Information). It provides insights into the growing use of memes, humour and Facebook live video streaming to spread disinformation in a manner that can escape detection and relies increasingly on organic dissemination.

Second, it presents research conducted by Lebanon's Maharat Foundation about the electoral campaign before the parliamentary elections of May 2022, which shows the high level of negative content in online campaigns, as well as a significant level of hate speech directed at women politicians.

Third, Al Hayat Rased, from Jordan, presents its research the on online abuse and harassment of supporters of the Children's Rights Act passed on 19 September 2022, with a special focus on gender-based hate speech, which shows the frequent use of religious references in hate speech.

Regional Context, Trends, and Findings

1. Regional context

Political instability in the MENA region is reflected in a distorted and unbalanced environment on social media platforms, where it becomes unsafe to engage in political discussions. The lack of independent media, of large-scale fact-checking, and of online literacy initiatives contributes to the uncontrolled spread of disinformation. In such a scenario, extreme viewpoints may be amplified, while more moderate perspectives are confined to the background, leading to a toxic and polarising atmosphere that makes it difficult to have constructive and meaningful conversations on these platforms. The lack of trust in traditional media can contribute to the spread of misinformation and disinformation, which can further distort the digital debate space. In such an environment, it becomes challenging to access reliable and accurate information, leading to a lack of informed discussions.

In the countries covered by our regional project, as elsewhere, some political actors use political disinformation on social media platforms to manipulate public opinion. This pattern, evident in the methods employed by networks associated with these actors, extends beyond electoral contexts, and permeates all online politically related discourse. Such efforts produce an environment in which harmful narratives can spread, affecting the reputation of individual political actors or political parties. The proliferation of these narratives on social media platforms results in the emergence of authentic-harmful-behaviours, such as hate speech propagated by the followers of these political actors.

Social media platforms, and especially Meta, the parent company of Facebook, Instagram and WhatsApp, are struggling with Arabic content moderation, a major problem in the MENA region, where the language is spoken with a great diversity of dialects. The continuous reliance on automated content moderation and machine translation, as opposed to increasing the number of human moderators from different countries in the region, has often led to missing the nuances of the context of speech and, as a result, to either underenforcement or overenforcement of content moderation policies, thus contributing either to harm or to censorship.

While this is neither new nor specific to the MENA region, the report identifies a change from past behaviour; evolving from amplifying messages through robots, which can be easily monitored and handled by social media platforms, there seems now to be a stronger reliance on visual messages (memes and videos) and live videos, which lead to organic dissemination on a much wider scale.

2. Regional Trends

2.1. First regional trend: Online gender-based violence

Women active in the public sphere are often targeted and subjected to harmful behaviours on social media platforms. This is particularly prevalent during times of elections and crises, when the most harmful and aggressive behaviours tend to spread.

In Lebanon, Words Matter partner **Maharat** detected hate speech on Facebook pages and groups targeting political actors and candidates. These pages and groups were selected based on whether they are promoting opposing or revolutionary actors [4]. the classification of the monitored content included abusive speech based on gender and gender identity, with expressions referring to sexual acts.

Forty-three per cent of the social media accounts of 100 active women candidates monitored showed various forms of online violence against women. This was mostly in the form of psychological violence (bullying, abuse, prejudice and threats, as well as intimidation directed at gender or social stereotyping) (85 per cent), with the remainder shared equally between sexual violence (directing phrases or content of a sexual nature at a candidate) (6 per cent) and comments on their age and appearance (6 per cent).

Figure 1: The breakdown of types of online violence against women during the 2022 Lebanese parliamentary elections, retrieved from 43 Twitter and Facebook accounts of women candidates

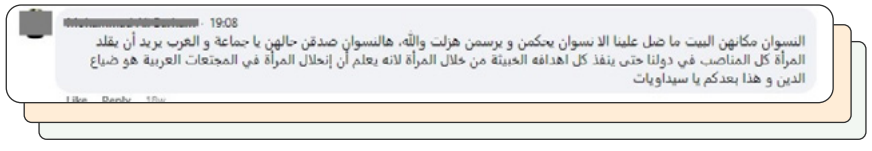
[4] Revolutionary actors is a term used to define the individuals and organisations who supported and participated in the revolution of 17 October 2019.

Maharat studied the case of the campaign against the MP Halima Kaakour. There was an intense debate between her and the Speaker of the House of Representatives, Nabih Berri, in which she objected to the way he managed the session, not giving the speech to the representatives in the speaking order. Berri addressed her by saying “Sit down, wait for the others, and be quiet”. Kaakour reacted by saying that the speaker was exhibiting patriarchal behaviour. These words sparked a social media discussion that included both posts offensive to her and in solidarity with her.

Figure 2: Screenshots from tweets and replies using the hashtag of the name of the MP Halima Kaarour attacking (on the left) or supporting her (on the right)

The attack targeted at female MPs in Lebanon prompted Amnesty International to send a letter to the Lebanese parliament and Berri, denouncing these practices and calling on the legislature to "take responsibility to end this abuse of women and stop encouraging a hostile environment against women."

In Jordan, Words Matter partner **Al Hayat Rased** monitored hate speech and misinformation campaigns around the draft Children Rights Act introduced to the parliament and examined from 20 July 2022 to 29 September 2022, date of its approval. They found that online abuse against those who supported the draft law was extremely gendered. Men were attacked in relation to their social status, deriding their roles in their clans and in society, and questioning their intentions. In contrast, women were attacked on many levels: using religious discourse and social discourse, for having foreign family origins, or by being characterised by a stereotypical role of women, women's rights, and their place in political life.



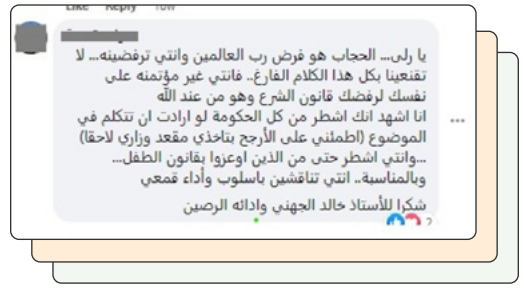


Figure 3: Screenshots of comments extracted from a sample of 346 posts on Facebook using social stereotyping and religious discourse against former MP Rola Al-Hroub.

Also In Jordan, Words Matter partner **JOSA** confirmed the results of Al-Hayat in their initial research. They found that gender-based online violence has taken on a more religious character, with commentators wrapping their offensive comments with a religious cover to justify their hate and anger, perhaps most notably in the case of Jordanian footballer Aya Majali, who made her debut in the UEFA Women's Champions League. Most comments denounced her wearing short shorts, even though she professes Islam. However, JOSA have not to date observed any content that calls for physical abuse. Most comments could be classified as:

1. Accusing women of engaging in Western agendas aimed at dismantling Arab and Islamic society, and at distorting its moral system.
2. Bullying and insulting women
3. Gender stereotyping (e.g., women belong in the home, kitchen, etc.).

In Sudan, **SUDIA** documented a repeated pattern of misogyny and gender-based violence on social media platforms, as well as on the ground, during public political demonstrations. Social media pages frequently criticise and mock women's participation in political movements and demonstrations, often posting pictures of them accompanied by sarcastic comments about their appearance and how they are dressed. This toxic environment encourages the public to engage in harmful discourse towards women, as evidenced by the high number of comments on these posts that could be classified as gender-based hate speech.

A screenshot capturing posts mocking the physical appearances of women participating in demonstrations.

A screenshot capturing posts mocking the physical appearances of women participating in demonstrations.

These women, who were already subjected to physical harassment and assault at the demonstrations, faced further online violence. The documented testimonies of physical assault on women during their participation in demonstrations, as well as the harassment and defamation they face on social media platforms, create a challenging environment for women to fully participate in political life. This behaviour limits their presence in the political arena and makes it difficult for them to fully integrate into this sphere.

2.2. Second regional trend: Religious-based hate speech.

In Lebanon, Maharat documented the use of shallow or cheapfakes, using audio-video manipulation to change backgrounds, manipulate audio, and add photoshopped logos of political parties, as in the case of the "Nazir Habashi" video, to spread hateful content on the basis of religion. Nazir Habashi, who claims [5] to be a Shia religious leader in Lebanon and a member of Hezbollah, appears in a video in which the background has been manipulated to warn against the Lebanese Forces Party, which, according to the video, poses a threat to other religious groups and their social practices.

Maharat also found that the Lebanese MPs who adopted a civil marriage law have been subjected to a religious hate campaign on social media, led by Sunni Muslim Sheikhs.[6]

In Jordan, **Alhayat-Rased** documented the use of negative religious discourse against women at all levels, regardless of whether they were wearing the Hijab or not. This was manifested in the criticism of former MP Rola Al-Hroub for not wearing the hijab, and for her comments about law and Sharia.[7]

[5] News Article by alnahda news, 02 May 2022, available in Arabic here : حزب الله يستنكر ما ورد في فيديو نظير حبشي ويأسف لاستغلال البعض تصريحاته المشبوهة (alnahdanews.com)

[6] Examples from two Sheikhs attacking MPs who voted in favour of the civil marriage law:
- Hassan Moraib shared video on Twitter – 2022-05-22, available here
- Al Hussein video available here

[7] Loosely translated from the comment "إنتِ سافرة وبتكلمي عن الشريعة", the word often carries a negative connotation for women not wearing Hijab and qualifies them as "indecent".

2.3. Third regional trend: The use of emotionally induced discourse in political manipulation.

In Tunisia, disinformation content, involving propaganda, conspiracy and political manipulation, bears a strong emotional charge, so as to amplify its dissemination. Triggering strong emotions is a great tool for making mis- or -disinformation go viral, possibly because strong sentiments can reduce the ability to analyse information objectively. In the below evidence, a collage of picture of political leaders from different background with the mark 'cancel' and a narrative of encouraging non-civil participation during the referendum and inflammatory speech against political parties.

Figure 4: An example of Tunisian content involving a manipulative tactic that plays on users' emotions.

In Lebanon, during the campaign for parliamentary elections, it was found that 49.5 per cent of the online discourse was based on the manipulation of sentiments, while posts by candidates and parties to promote their political platforms came a distant second, at 21.3 per cent.

Traditional political forces have relied more on emotional rhetoric to strengthen partisan and sectarian affiliation. The propaganda of political parties continues to bring up the memory of the 1975 to 1990 Lebanese Civil War, to fuel strife and sectarian and political conflict, and to evoke conspiracy theories to demonise political opponents and damage their reputations and credibility.

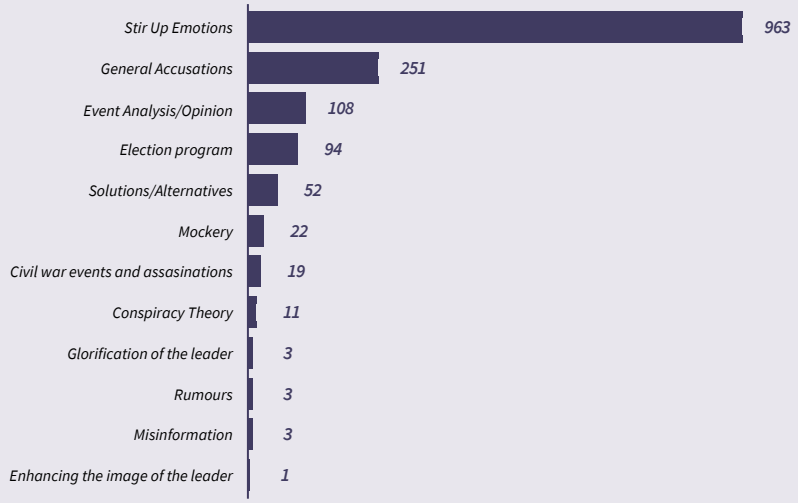


Figure 5: The breakdown by theme of 1,943 Tweets and posts by 107 candidates and politicians during the electoral campaign

2.4. Fourth regional trend: Coordinated disinformation campaigns on social media during major national political events.

In Tunisia, Lab'TRACK monitored coordinated political live broadcasts on Facebook as a space for spreading disinformation. Some political actors on social media use the "Live" function of Facebook and broadcast the stream simultaneously on several pages, allowing them to reach many followers. These live broadcasts have been documented to contain hate speech and misleading information. The use of this feature also makes it possible to interact directly with specific communities.

3. Regional tactics used in spreading hate speech and online manipulation

3.1. The weaponisation of political memes

One of the main techniques and strategies for disseminating harmful content is the use of satire through "memes" and humour. In Tunisia, Lab'TRACK documented satire on

Facebook pages using memes to spread harmful content. One of the regular targets on Tunisian Facebook is Abir Moussi, president of the Free Constitutional Party (Parti Destourien Libre) and member of the Tunisian Parliament, prior to its dissolution, on 25 July 2021.

Figure 6: Screenshot showing the combined use of satire and disinformation in the same post as a tactic to discredit Kais Said's opponents.

3.2. Coordinated Live videos

According to Facebook,[8] users comment 10 times more on live videos[9] than on pre-recorded videos. Live videos, notified in advance to page and group followers, have a higher organic reach and generate greater engagement than pre-recorded videos. They are often broadcast simultaneously on multiple pages making it hard to detect though platform content moderation filters An example of this was observed in Tunisia during the 25 July 2022 referendum.

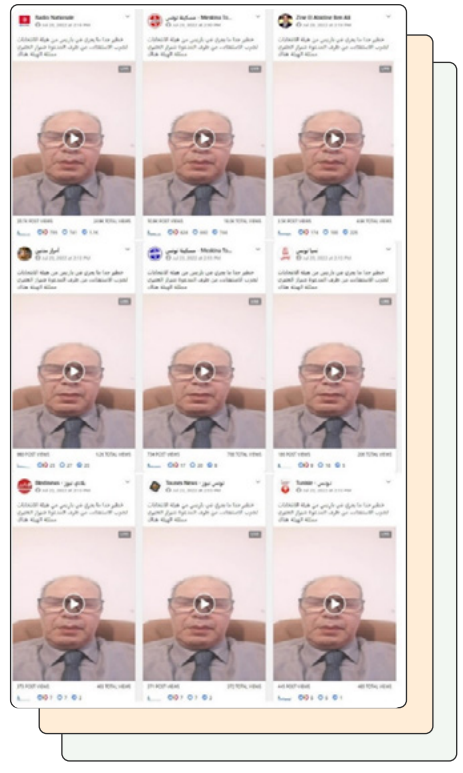


Figure 7: Screenshots showing the coordination between different pages live streaming videos simultaneously.

[8] Simo, "Introducing New Ways to Create, Share and Discover Live Video on Facebook", Facebook, April 2016.

[9] Some of these videos, presented as live, are in fact pre-recorded videos.

Spreading misinformation and hate speech through live videos makes it harder to detect with automatic content moderation – even when flagged by Facebook users, the reaction from Facebook is not instantaneous, allowing the message to spread.

3.3. Comments contain more hateful content

Words Matter's partners noticed a higher frequency of harmful content (hate speech) in comments than in posts. Maharat noted that comments on Facebook were used heavily in gendered and religious hate speech campaigns in Lebanon. Al-Hayat, in Jordan, documented a similar partner in the comments section on Facebook on the topic of the draft Children's Rights Act, which they had already noticed when observing the previous municipal elections (see Words Matter's first report).[10]

3.4. Activation of fake accounts during elections and democratic transition

In Lebanon, political actors have used Twitter for complex amplification, through targeted campaigns aimed at influencing political opponents. Maharat documented the behaviour of fake accounts created to amplify specific political messages and participate in defamation campaigns against politicians. These accounts were only active during the parliamentary elections. In the example below, a Twitter account is sharing harmful rumours about candidates, including of their withdrawal from elections or accusations of bribery. The account was active between May 10 and 14, and published many of these memes.

Figure 8: An extract from a fake account created on 16 January 2022 on Twitter, with 1,033 followers and no following.

[10] Online Public Discourse in MENA: Disinformation and Hate Speech During the 2022 Lebanese and Jordanian Elections", Democracy Reporting International, 28 September 2022.

4. Regional Recommendations

The recommendations below are based on evidence collected by the Words Matter Network. They are meant to enhance information integrity on social media platforms.

For political parties and movements

- Develop an internal code of conduct to prevent members from engaging in hate speech and disinformation on social media.

For traditional media:

- Engage with civil society and peer organisations in the country and the region to develop and commit to reporting standards that do not promote hate speech.

For the community of researchers in the MENA region,

- Improve collaboration between computational researchers, digital rights activists, sociologists, and other analysts, on a regional scale.
- Remain vigilant in relation to new forms of disinformation and to the use of AI in generating and spreading disinformation and hate speech through synthetic media and deepfakes.

For Social Media Companies

- Invest more in content moderation, by hiring local researchers from different parts of the region to perform human content moderation. Enhance the language modules of algorithms to detect hate speech in Arabic and its local dialects in the region.
- Collaborate with researchers in the MENA region to develop hate speech lexicons in different dialects to detect different forms of harmful content in the region.
- Enhance their reporting mechanism and create a balance between machine and human responses to harmful content.
- Work closely with civil society and electoral bodies to introduce and explain their community guidelines and content moderation policies.
- Be vigilant with regard to new tactics of bad actors, for example, to the spread of hate speech in the comments section on Facebook, the use of text overlay in images (memes), leaks, and doxing the private information of political candidates. Develop further technical capabilities to monitor harmful content in Arabic shared via live videos.
- Expand their networking programmes and partnerships with civil society initiatives concerned with democracy and promoting a digital environment that guarantees human rights. Support independent media platforms promoting safe spaces for women active in the political field, and digital media literacy initiatives in general.

- Lift restrictions on new users of monitoring application tools, such as CrowdTangle, and facilitate access to information for CSOs and research and academic institutions.[11]

In addition to the above, Words Matter firmly believes that social media platforms should cooperate with relevant organisations and CSOs to come up with policies that contribute to reducing gender-based digital violence in the region. The quest to provide a safe digital environment for women has, so far, been limited to the national level, with scarce (if any) joint work with those platforms. The responsibility to protect vulnerable groups in digital spaces lies not only with national governments and CSOs, but also with social media companies and other tech actors, on a regional or global level.

Civil Society

- Strengthen regional collaboration and engage routinely with different stakeholders, including governmental bodies, legislators, specialised independent authorities and tech companies.
- Develop special programmes to monitor, document, and analyse harmful content on social networking sites.
- Launch more initiatives for information verification, networking, and exchange of experiences.
- Organise national workshops on legislative amendments or improved and transparent enforcement to better deal with the challenge of hate speech, disinformation, and false news. This should include the participation of judges dealing with these legal issues.
- Form specialised, multi-disciplinary bodies to better demand and monitor corporate accountability from tech companies.
- Create helplines to support victims of online gender-based violence and assist them in reporting harmful content to social media companies, to advise them on digital security and presence, and to provide them with resources to support them during and after online campaigns targeting them.
- Promote digital literacy among social media users.

[11] Since January 2022, Meta has stopped the right of new users to use the CrowdTangle tool to search social media.

Research Limitations at the Regional Level

Words Matter partners faced operational challenges during the period of the project, especially in Sudan and Lebanon, where they suffered from power cuts, demonstrations, and political troubles, which negatively affected their work. Challenges specific to the monitoring of online abuse included:

1. The fact that most of the tools used in collecting data, such as Twitter AP and CrowdTangle, are not suitable for social media monitoring in Arabic, because of the variations between dialects and standard Arabic. They are even less suitable for other local languages spoken in the MENA region.
2. Difficulties choosing specific locations on CrowdTangle.
3. The news that Meta was possibly planning to stop supporting CrowdTangle, hampering the work of social media researchers and fact-checkers who are building their methodology to research and collect data on CrowdTangle.
4. Some technical issues with ExportComment tools, requiring the sending of multiple requests to the tool, delaying data collection
5. Different national contexts and different issues surveyed, preventing Words Matter network partners from using a unified methodology to monitor hate speech and disinformation. This made support to partners more time-consuming and cross learning less fruitful.
6. Online abuse and manipulation were documented, but the environment, whether legal, technical, or societal, that enables it remains to be properly explored for better and more specific suggestions for improvement. This will be a priority of the next report.

Country Case Studies

1. July 25th, 2022, Tunisian Referendum

1.1. Context

The post-revolution period in Tunisia, from 2011 till 2021, opened the doors for people to participate in the public and political spheres. This period was marked by the emergence of political participation at different levels – local, regional and national. During this period, Tunisian citizens were called on to participate in four elections to elect their presidents, their deputies and their municipal councils. On 25 July 25 2021, President Kais Said suspended (and later dissolved) the parliament, and announced he was assuming emergency powers until a new constitution could be put in place. He submitted a draft constitution to a referendum (the first since the revolution) on 25 July 2022, which resulted in its adoption. This referendum was the first time that citizens had been called to vote since Said's decree centralising all powers in his own hands. The constitution itself was drafted by a small committee and largely amended by the president. There was virtually no attempt at transparency or at explaining the changes and issues in the new text, which turned the referendum into a vote for or against Said, rather than on the merits of the constitution itself.

1.2. Scope of monitoring

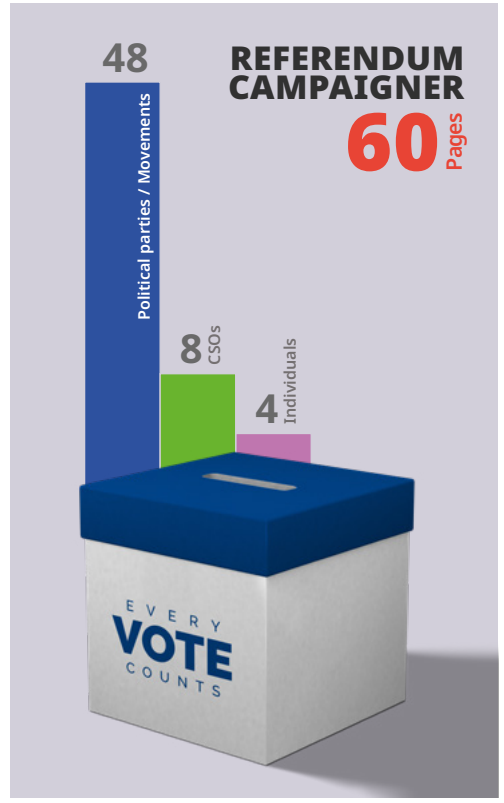
Lab'TRACK sought to understand the disinformation tactics that could be used during the referendum. The team monitored the most influential Facebook pages and groups aiming to influence, either positively or negatively, public opinion around the referendum.

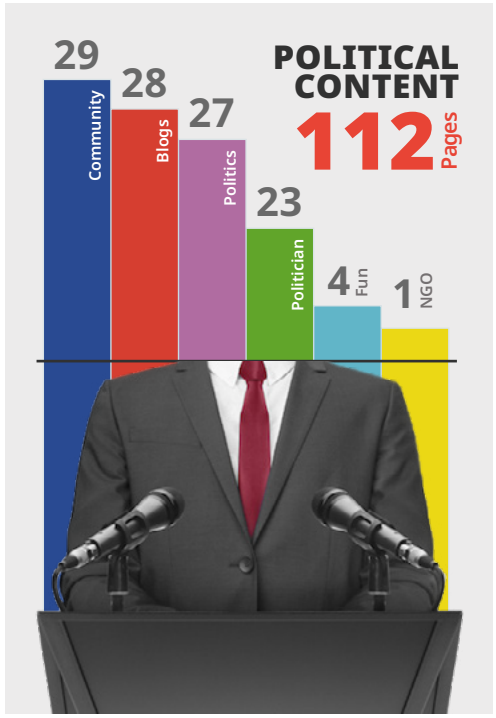
The monitoring focused on tracking misleading behaviour and harmful content aimed at affecting campaigns around the referendum, investigating new misleading tactics and dissemination of disinformation.

1.3. Methodology

Lab'TRACK created a list of Facebook pages and groups to be monitored. This list included media pages, pages belonging to individuals or groups campaigning for or against the adoption of the new constitution, and pages that regularly publish political content or engage in political discussions. The monitoring list comprised 270 Facebook pages: 112 pages that regularly publish political content,
96 media pages, 60 pages belonging to referendum campaigners, and
2 belonging to state entities. The monitoring list also covered 26 public groups.

The figure below provides a breakdown the different monitored pages and groups.





1.4. Data analysis

The data collected for this study was obtained by using CrowdTangle, an official Meta tool that enabled Lab'TRACK to collect content and publications from Facebook. Lab'TRACK also used the Factalyzer app, which assisted the team in annotating the content and manually verifying the aggregated data.

During the monitoring period, from May 25 to July 28 2022, the Lab'TRACK team managed a monitoring process that involved analysing a substantial number of Facebook posts, approximately 350,000. In order to understand which posts were most engaging, the team focused their analysis on the 6,500 posts generating the highest level of interaction.

1.5. Data classification

Lab'TRACK classified the posts into four categories:

- Political Content
- Disinformation content
- Violent or Hate Speech
- Content about female political figures

Total monitored content	Disinformation	Hate-Speech	Against Women	Political Content
6500 Posts	472 Posts	552 Posts	151 Posts	34 Posts

1.6. Misleading behaviours and tactics

1.6.1. Coordinated Behaviour

Lab'Track identified the use of coordinated political live broadcasts on Facebook as a means of spreading disinformation. Certain political actors on social media use the Facebook Live feature, broadcasting the stream simultaneously on multiple pages to reach a large number of followers, thus benefitting from Facebook algorithms. These live broadcasts were documented to contain both hate speech and misleading information. This tactic not only allows for the dissemination of false or harmful information, but also enables direct interaction with specific communities.

In addition to this, some of these political actors were observed presenting viewers with misleading contexts, such as appearing in official attire with the flag of Tunisia as a backdrop and discussing official updates, information and political analysis. This formal presentation might have misled viewers, creating confusion and leading them to believe that the speaker was a government official, contributing to the spread of false or harmful information.

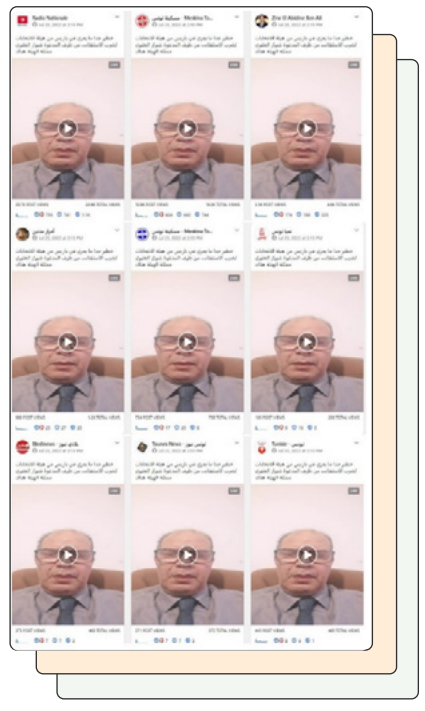


Figure 9: Screenshots showing the coordination between different pages livestreaming a person who is criticising the ISIE[12] office in Paris.

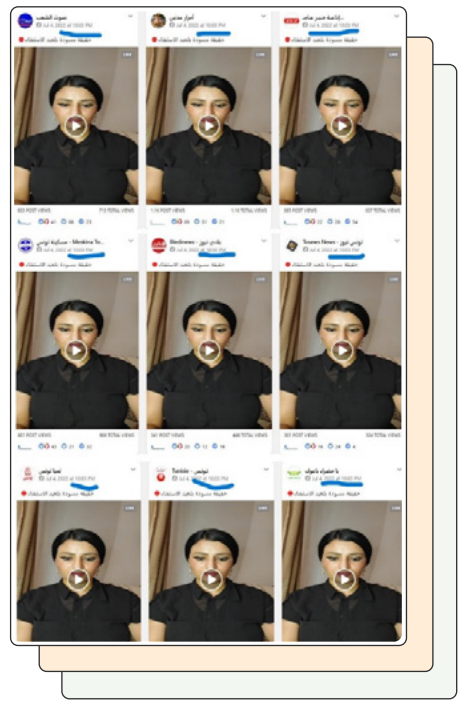


Figure 10: A figure showing different pages sharing a live video simultaneously of a person claiming to explain the truth about the text of the constitution text drafted by the Sadok Belaid commission.

[12] Instance Supérieure Indépendante pour les Élections (Tunisia's Electoral Monitoring Board).

The screenshot above demonstrates the existence of another feature of manipulation, consisting of publishing pre-recorded videos as "live" in the same time to amplify specific political messages to affect the turn out and results of the referendum .

It is important to highlight the difficulty of using automated moderation for the content of live videos, especially if the content is in Arabic or the live videos are very long.

1.6.2. Meme as a political manipulation tactic

The use of memes in the Tunisian political context has been found to have a significant impact on the dissemination of dis/misinformation. Lab'Track observed that the owners of some pages use satire, caricatures and memes to discredit their opponents. Observation suggests the existence of a link between the use of satire and humour and the spread of disinformation and political manipulation. The use of satire and humour can evade detection, as the line between humour and harmful content might be very thin, and can be particularly tricky to identify for an audience that is not familiar with the context.

The emotional effects that may be generated by the use of humour, such as amusement, joy, and shock, may explain why they facilitate the dissemination of disinformation. When an individual makes another person laugh, they are often perceived as more sympathetic, and manipulative content may be shared more readily without questioning its veracity.

Indeed, according to Lab'Track observations, satirical content – and "memes", in particular – can slip through the cracks, generating dissemination mechanisms and numerous reactions.

As per Figure 6 above, the meme is mocking Abir Moussa's (Head of political party opponent to the referendum) calling the leadership of the Tunisian Trade Union not to participate in the referendum, while the trade union president is showing his finger with the voting ink as a proof for voting with a disguised use for a swearword. The meme links two images with initially no direct relationship.

2. May 15th, 2022, Lebanese Parliamentary elections

2.1. Context

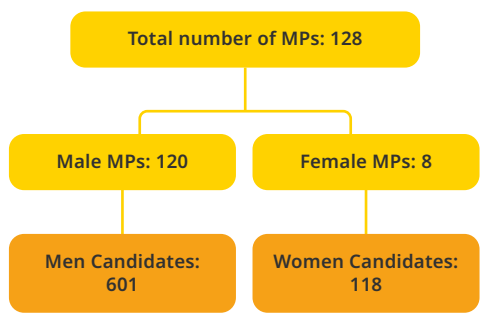
The parliamentary elections of May 2022 took place in an atmosphere of scepticism and amid accusations of bribery and candidates exceeding financial spending limits. In the absence of oversight by the Election Supervisory Authority, whose chairman acknowledged in several media statements the Commission's inability to control bribery and spending that takes place outside of the framework of legal regulation, such as through the disbursement of cash funds that are not documented, following the collapse of the banking sector and the restriction of bank withdrawals. Candidates complained about this

phenomenon, and called on the supervisory body, the security forces, and the judiciary to perform their related duties.

The campaigns of the opposition and revolutionary forces against the traditional parties and forces of the political system contributed to increased votes from the Lebanese diaspora, in favour of the opposition and revolution forces, and led to the victory of deputies from outside the traditional party alignments. This was unlike the 2018 elections, as civil society candidates at that time won only 2,379 of 46,799 votes, and the percentage of expatriate voting at the time did not exceed 2.5 per cent of all voters.

More than one observer, especially the European Union Election Observation Mission (in its report), spoke of "practices" of vote-buying that negatively affected citizens' freedom of choice and led to a lack of equal opportunities.

The Lebanese parliamentary elections, the final phase of which took on 15 May 2022, resulted in the victory of 13 deputies from emerging political forces, who entered the parliament at the expense of some political figures from the traditional political forces.



2.2. Methodology

This study aims to shed light on the political discourse and manipulation campaigns during a monitoring period that covered the period from 1 April to 15 May (election day), during the post-election monitoring period running up to the announcement of the results on 16 and 17 May, and through to the end of President Michel Aoun's term, on 31 November. This included:

- Monitoring speech on social media questioning the integrity of the electoral process during election day and after the close of polling stations, and then as the results were announced and while appeals were being submitted to the Constitutional Council.

- Monitoring the disinformation and manipulation campaigns (rumours, false news) that affected the 13 MPs who entered the Lebanese Parliament as representatives of the revolution.

- Monitoring the hate and abuse campaigns that all MPs were subjected to following the announcement of their victory and their official involvement in the parliament.

- Monitoring the political discourse towards women parliamentarians from the Forces of Change.

Monitoring Sample of Political actors:

A group of political actors, consisting of politicians, candidates, and influencers, comprising 132 personalities, 20 per cent of whom were women, were monitored on social media in April and May, to identify the pillars of the political discourse and the topics of that discourse.

The sample included 107 politicians and candidates representing various traditional and emerging political forces and the geographical and demographic divisions of the 15 constituencies, as well as 25 influential figures with multiple partisan and political tendencies. Women made up 20 per cent of the total actors monitored and 16.43 per cent of the total number of candidates.

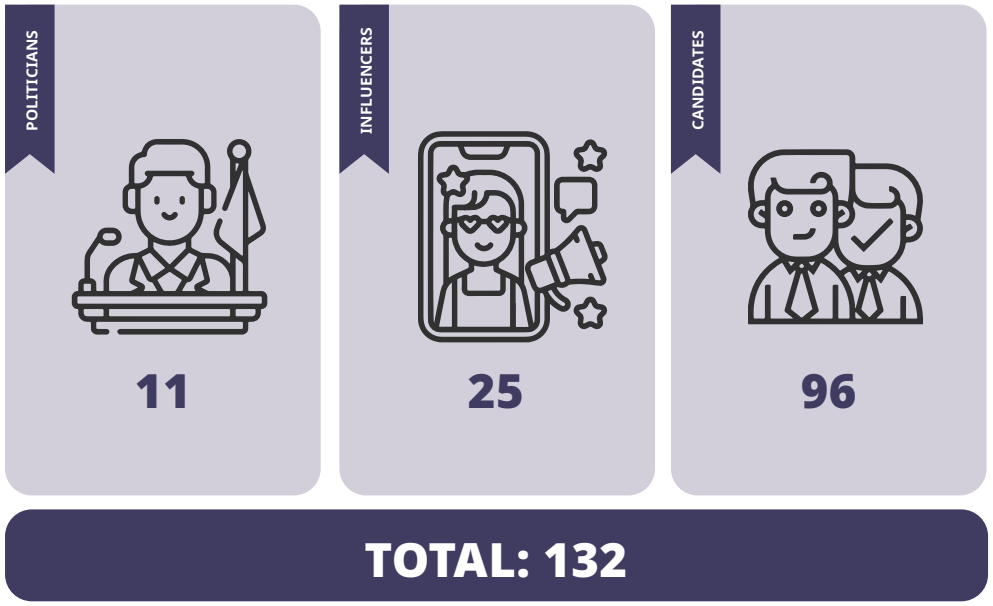


Figure 11: Distribution of monitoring sample

The volume of posts and tweets monitored on Twitter and Facebook from April 1 to May 15 totaled 2,628, with the majority of these being tweets, as most of the actors' accounts were monitored on Twitter as their primary platform. As for the actors who did not have accounts on Twitter or who rely on Facebook as a means of communicating with the public, they were monitored on Facebook.

2.3. Data Analysis

2.3.1. Political propaganda

The distribution of the type of political propaganda disseminated by candidates and politicians on social media, Facebook, and Twitter, according to the monitored sample of 1,943 posts/tweets from 1 April to 15 May, was as follows

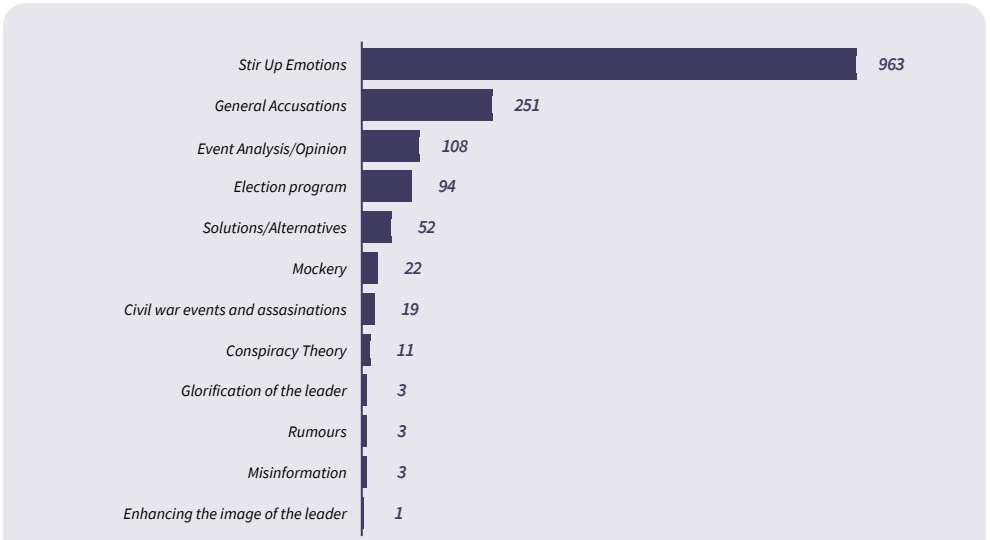


Figure 12: Breakdown of 1,943 Tweets and posts by 107 candidates and politicians during the electoral campaign

Stirring up emotions refers to discourse based on attacking specific political opponents and accusing them or posting negative and stereotypical images of them to damage their reputation and distort their political image.

Pillars of political propaganda:

The share of political propaganda posts based on emotional speech was 49.5 per cent. In second place was self-promotion, at 21.3 per cent. The total political discourse related to electoral programmes and proposing solutions and alternatives was just 7.5 per cent.

Topics of political discourse on corruption, development, and human rights:

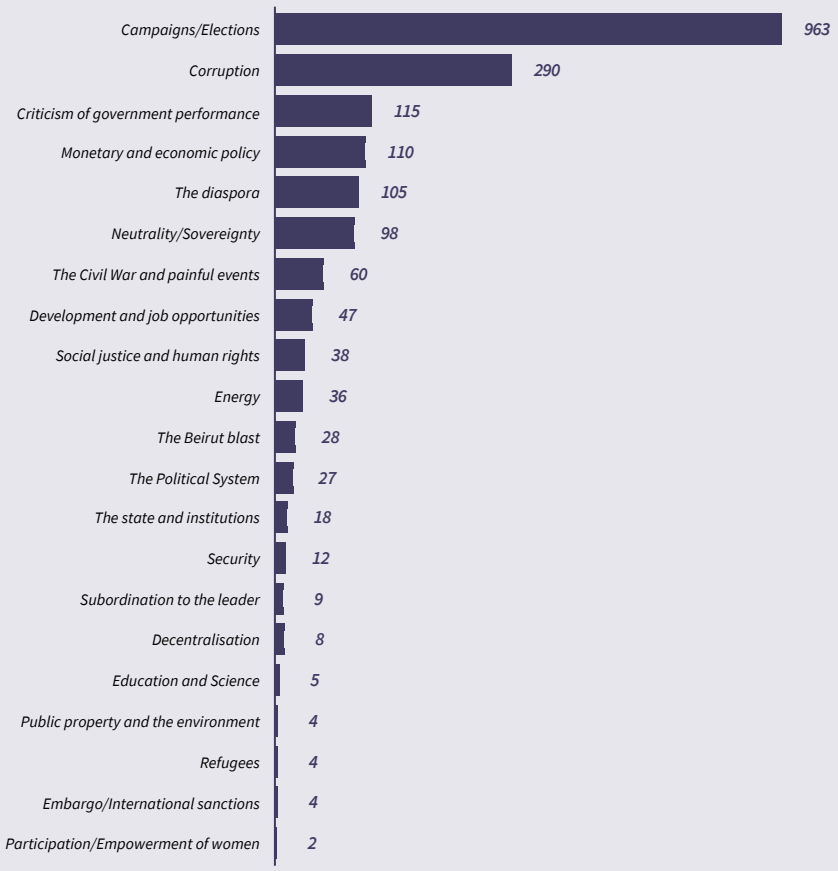


Figure 13: Breakdown of topics touched on by 107 politicians and candidates from a sample of ,1943 posts/tweets during the electoral campaign.

The issue of corruption[13] constituted a basic theme for political propaganda for political actors, appearing in 15 per cent of the posts; the share of political discourse on development issues and employment opportunities was 2.41 per cent; and topics related to social justice and human rights, and to energy-related topics each appeared in 2 per cent of the posts.

[13] For a full analysis of most discussed topics, see "Political Propaganda and Information Manipulation on Social Media During the Lebanese Parliamentary Elections", Maharat Foundation, 2022, p. 27 and seq.

The negative trend of public discourse during the election campaign

The rhetoric of candidates and politicians during the election campaigns was largely negative, with an inflammatory character that exacerbates conflicts. The share of negative speech amounted to 71 per cent of political discourse analysed, as shown in the figure above. Negative and high-pitched rhetoric on social media pages was also accompanied by violent rhetoric promoting hatred.

2.3.2. Rumours and misinformation during election campaigns

- A wide range of Facebook pages were involved in political campaigns and the promotion of political actors. (Promotion of opposition revolutionary and transformative parties and forces).
- Forty-two of the Facebook pages and groups analysed contributed to the spread of rumours, as 79 rumours and six pieces of false news were posted between 1 April and 15 May.
- Many fake accounts, created for specific purposes and periods, were identified on the Twitter platform.
- Some of the fake accounts identified were used to launch campaigns to attract votes, such as "your boycott serves them", or to attack candidates and their reputations, such as "the story of Paula Yacoubiyan list" and "don't be like Jad".

Figure 14: Example of fake account identified by Maharat

Fake accounts were used to spread rumours about candidates, aimed at discrediting them, or spreading news that they had withdrawn, or accusing them of bribery and vote buying. These accounts were only active during the parliamentary elections. In the example below, a Twitter account shared harmful rumours about candidates or about their supposed withdrawal from the elections, as well as accusations of bribery. This account was active between from 10 to 14 May and published many of these memes.

Figure 15: Example of a fake account on Twitter identified by Maharat

On 22 April 2022, at least 51 Twitter accounts participated in a smear campaign against candidate Jad Ghosn, using the hashtag #ما_تكون_متل_جاد, accompanied by a video sullying his reputation, using false information aimed at destroying his credibility.

Figure 16: Examples of tweets attacking Jad Ghosn

One journalist described the campaign against Ghosn as an "intense childish Twitter campaign". He said the campaigners understood to what extent he was a competitive candidate, asking Ghosn to ignore them and follow his campaign.

Activity on Twitter during the election campaigns:

Political actors used Twitter for complex amplification through targeted campaigns aimed to undermine political opponents. .

Between 1 April and 31 May, 134 trends were monitored on Twitter, 69 per cent of which were motivated by partisan or politically oriented actors.

Among the most active political actors in relation to campaign-related hashtags/trends, the Lebanese Forces activists came first (25 hashtags), followed by supporters of the Free Patriotic Movement (20 hashtags) and Hezbollah (18 hashtags). Following the spread of a manipulated video purporting to show Sheikh Nazir al-Jishi, a religious leader, attacking Hezbollah for being hostile to some components of Lebanese society, specifically Christians, the hashtag: their culture is death, ours is life was used by opponents of Hezbollah.

Trending on Twitter with the support of Hezbollah's political opponents and identification with the Lebanese Forces party, this hashtag was widely circulated according to the following deployment map:

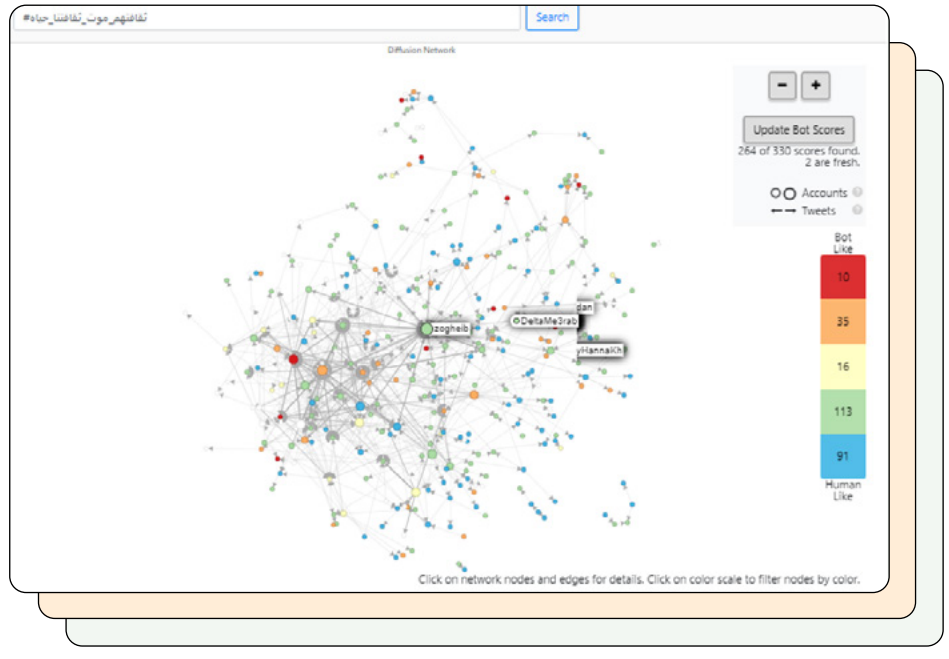
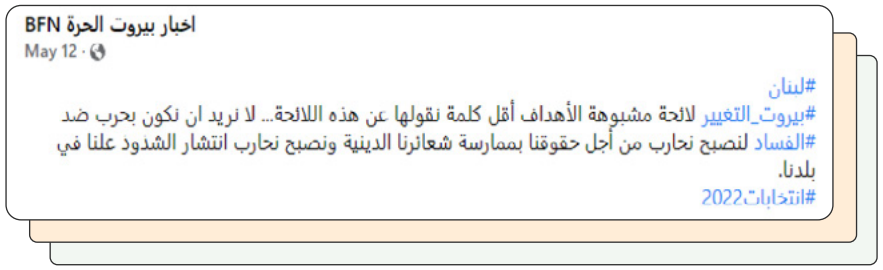


Figure 17: Development map of the hashtag "their culture is death, ours is life was used by opponents of Hezbollah

2.3.3. Promoting violence and hatred during election campaigns

Some accounts have been used to incite violence on religious, regional or factional grounds, according to political opinion, as well as campaigns making accusations of treason and apostasy.



Hate speech has emerged in social media campaigns, with some candidates being subjected to personal campaigns related to their sexual identity, national origin, belief or opinion.

Hate speech against Omar Harfoush, a Sunni candidate from Tripoli, was identified on Facebook, with abusive speech based on gender and gender identity. Exposure to it in terms that include sexual acts. Like a "fun candidate"

Photos and videos circulated that included stereotypical expressions and accusations, as illustrated below:

Figure 18: Screenshots from tweets attacking the candidate Omar Harfoush, calling him a "fun candidate"

Candidate Wadah al-Sadiq has also been subjected to a hate campaign related to his non-Lebanese origins. In this tweet below, the user shames the candidate for his Armenian origins.

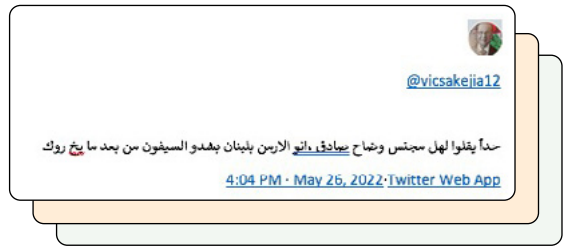


Figure 19: Screenshots from a tweet in Arabic attacking candidate Sadiq for having Armenian origins.

2.3.4. Campaign of rumours, misinformation, and false news against the MPs of Change forces

Rumours were spread about the Change forces MPs, with the aim of offending them and tarnishing their image, and media pages participated in these campaigns.

Information was circulated that MP Mark Daou called for more United States sanctions against Lebanon. This information was fact-checked, and found to be incorrect.

The Twitter campaign was led by journalists, media websites and pro-resistance party activists,[14] Hezbollah, and the Free Patriotic Movement, demanding that his parliamentary immunity be revoked and that he be tried for crimes of dealing with and inciting hostile parties against Lebanon.

Figure 20: An example of tweets attacking an MP from the Change forces, Mark Daou, based on false claims that he called for more foreign sanctions against Lebanon

[14] A group of individuals and parties close to Hezbollah. After 2022 elections, they are still in Hezbollah camp against Geagea.

2.3.5. Campaigns of religious discourse against Change forces MPs:

Change forces MPs have been subjected to a hate campaign on social media, fuelled by clerics opposed to their work for the adoption of a law introducing civil marriage in parliament, after they appeared on 19 May in a panel discussion with Marcel Ghanem on Lebanon's MTV program "Sar al-Waqt."

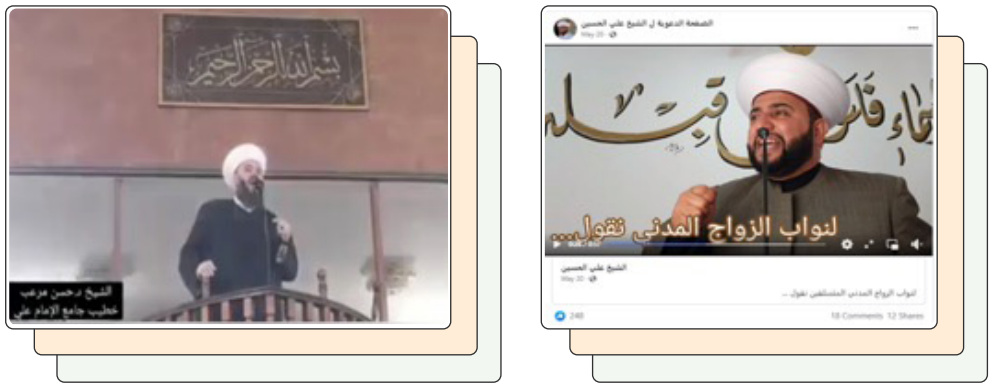


Figure 21: Screenshots from Sheikh Ali Hussein posts on his Facebook account

This campaign was heightened by posts by two Sheikhs on social media.

Sheikh Ali Al-Hussein stated in a sermon published on his personal Facebook account that "three MPs in Beirut, male and female", (whom he refrained from mentioning by name)," who succeeded in the votes of Muslims, ... Yesterday, they proposed civil marriage and announced their consent. There is a fatwa by the former Mufti Qabbani that whoever marries in a civil ceremony, rather than religious is considered an infidel and an apostate. Yesterday they declared war on Islam and declared war on God, and we will declare war on you."

Sheikh Merheb addressed his speech to the Change forces MPs Ibrahim Al-Munimneh, Waddah Al-Sadiq and Halima Kaakour, whom he named by name, and said: "I say to those who call themselves changers... and everyone of their shape. If you do not repent to Allah and announce your return from this project, you are outside the religion of Allah and you are apostates from the religion of Islam, and we are innocent of you."

2.3.6. Violent speech against women working in politics

Violent speech against women candidates

Many women candidates were subjected to online violence on their social media accounts. In monitoring the accounts of 100 active women candidates on social media, it was found that about 43 per cent of them were subjected to some form of online violence.

An analysis of the form of violent discourse directed at women candidates through comments on their accounts and social media activities revealed that 86 per cent of violent responses and comments took the form of cyberbullying (abuse, ridicule), and about 6 per cent came in the form of violence based on appearance and age, as well as a similar percentage in the form of sexual harassment. The share of comments specifically biased against women, including gender stereotypes, the role of women in society, and the patterns that should be followed and imposed on them by society and the environment in which they live, was 1 per cent.

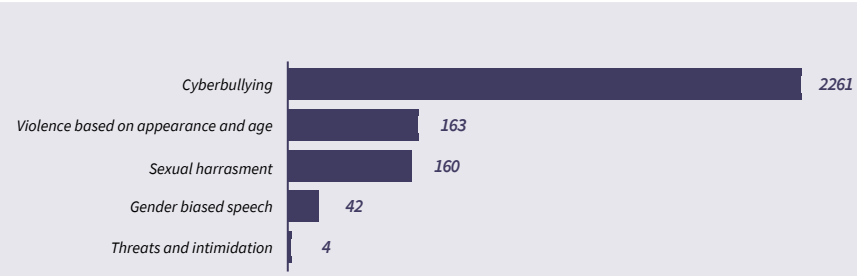


Figure 22: Classification of online violence against women from a sample of 2630 comments on 43 women candidates on Twitter and Facebook

Patriarchal discourse directed at women parliamentarians

Only 8 out of 118 women candidates won their races in the elections of 15 May 2022. Among them were three women who belong to the Change forces. With the start of the parliament's work, the misogynistic discourse moved from social networking sites to the corridors of the Lebanese parliament.

In this paragraph, we will review two case studies. The first relates to the sexual harassment and verbal harassment of **MP Cynthia Zarazir**, and the second to the harassment of **MP Halima Kaakour**, following her objection to the speaker's style of conducting sessions.

Sexual harassment and verbal harassment of MP Cynthia Zarazir

Zarazir stated in a side statement during a session of parliament session held on 26 July 2022 that she was harassed and verbally bullied and abused by her fellow MPs in the session, and mocked with epithets such as “cockroach” (her surname rhymes with the Arabic word for cockroach) and “starling”. She noted that she had been harassed since the first day she entered parliament, and posted an explanation on her Facebook page of the types of harassment she was facing. Social media users then circulated a hashtag with her name in Arabic letters #سينتيا_زرازيري to comment on her statement. Some stood in solidarity with her, others justified the harassment, and others yet attacked her. Among the accounts used, profiles of women who participated in the campaign against her were prominent.

Figure 23: Screenshots of tweets attacking or supporting MP Zarazir

MP Halima Kaakour accused of insulting a religious site and blasphemy

During a session of the House of Representatives, on July 26, 2022, a debate took place between the Speaker of the House of Representatives, Nabih Berri, and MP Halima Kaakour, who objected to the way Berri was managing the session and that he was not allowing MPs to speak according to the speaking order. Berri addressed her, telling her to "Sit down and wait for the end and be quiet.” Kaakour responded by rejecting this "patriarchal method"[15] of conducting the session. She stated in a video published on social media that Berri’s behaviour represented the moral denigration and depreciation of the role of women

parliamentarians, adding that "electronic armies" are behind these campaigns against them.

A controversy ultimately erupted that crossed over to social media, where a group of media pages on Facebook claimed that Kaakour’s words used the word “Patriarchal” to refer to the head of the Maronite Church, and not [to Berri]”.

The use of the term “patriarchal” by Kaakour to describe Berri's behaviour was objected to by MP Farid Heikal Al-Khazen. He requested that the sentence be deleted from the minutes of the parliament, a request that was accepted by the speaker, who issued a statement to the media about “harming the sacred”. The discussion deviated into a campaign to defend the patriarchal see and Bkerki (the head of the Maronite Church in Lebanon), placing the word “patriarchal” in a sectarian and religious framework unrelated to what had Kaakour meant.[16]

Kaakour was subjected to a hate campaign that included the use of terms such as "stinky”, “atheist”, “ISIS" and others, but also enjoyed the support of those showing solidarity with her, as the following screenshots show:

Figure 24: Examples from tweets using the hashtag “Halima Kaakour” (in Arabic letter) showing divergent reactions spreading hate speech against the mentioned MP

[15] Patriarchal discourse is based on a social organisation characterised by the supremacy of the male head (father) and the subordination of women and offspring to him, including legal subordination.

[16] The head of the Maronite Church is also called the patriarch, and the hate campaign was based on the confusion of the use of the term with a different meaning than was meant by Kaakour.

Amnesty International showed solidarity with women parliamentarians who are subjected to violence from their male counterparts.

2.4. Conclusions

Candidates' political discourse was not based on electoral programmes, as discourse related to electoral programmes and proposing solutions and alternatives amounted to only 5.7 per cent of the posts by the monitored political actors on social media.

The political discourse of political actors contributed to campaigns that increased disinformation and the intensity of conflicts on social media, as the monitoring of influencers and active pages showed that most of them used provocative speech, inciting emotions and spreading political propaganda associated with promoting the leader or making offensive statement about political opponents. Some campaigns started with a hashtag launched by the party leader. In addition, a large number of groups and pages contributed to the promotion and spread of baseless rumours.

The involvement of party supporters in misinformation campaigns, the manipulation of information, and harassment – reaching the level of hate speech in some cases – confirms the absence of an organisational framework that that stimulates transparency and integrity in the partisan work of political parties.

Out of 20 campaigns monitored whose hashtags trended on Twitter, seven were campaigns against MPs from the change forces between 18 May and 18 June. These campaigns covered different political instances related to political views, parliamentary activity or their reactions during the celebration of the victory.

MPs from the Change forces who adopted the civil marriage law have been subjected to a religious hate campaign on social media, led by clerics.

The patriarchal rhetoric and harassment of women parliamentarians in parliament by its speaker and some men MPs have been reflected on social media by hate campaigns and justifications for violence and bullying.

3. Children's Rights Act and online hate speech in Jordan

This section will focus on how a heated debate over the draft Children's Rights Act, which led to hate speech campaigns on social media led by various political and religious leaders. The research was conducted by Al Hayat-Rased, through social media monitoring from 20 July to 28 October 2022.

Online gender-based violence. The case of Children's Rights Act provides evidence of how far hate speech can reach to

attack women, using different tools to spread online gender-based violence, as well as an opportunity to identify more effective responses to protect women online. As such, interest in women's issues in technology has begun to increase in Jordan, like other countries in the world, which has resulted in the emergence of research initiatives that seek to study the status of women in public digital spaces. The national debate on the inclusion and empowerment of Jordanian women in the public and political spheres is also repeated from time to time, but to this day there has not been success in developing an integrated system to protect women from gender-based digital violence. Women remain one of the groups most vulnerable to digital violence and most affected by its consequences, and legislation has been unable to alter this situation, partly as a result of the fact that, to a large extent, the system does not consider the social and cultural reality of women in Jordanian society and the consequences of this reality.

For example, women in Jordan, according to our partner JOSA, tend not to file official reports of digital threats they face, due to the social consequences of doing so. The results of JOSA's surveys for a policy paper[17] over the past year confirm this, with most women surveyed reporting that the consequences of some forms of digital gender-based violence may even be being killed by a family member (often referred to as "honour killing"). Unfortunately, reporting digital violence is a complex, bureaucratic and insecure process for women, due to litigation procedures that require the complainant's personal presence and do not consider the sociocultural specificity of women in Jordan. It is also worth mentioning that, in a study on violence against women in the public and political spheres recently published by the National Commission for Women,[18] only 12 percent of the women participating in the study had turned to the judiciary in such contexts.

Further exacerbating the situation with gender-based violence is the fact that a large share of those working in the relevant official authorities do not have sufficient knowledge and training on gender concepts and related issues to deal effectively with such cases. According to a UNHCR study in East Amman, many women's service providers reported that they often did not know how to deal with cases of gender-based violence.[19]

All the above makes social media platforms in Jordan unfriendly spaces for women, given the harassment they may face on these platforms, especially if they are political or human rights activists.

While JOSA is working on developing an AI tool to detect hate speech against women on Twitter in Jordan, our second partner, Al Hayat-Rased, has examined the impact of the debate on the draft Children's Rights Act on the rise of hate speech, and particularly hate speech against women, as explained below.

---

[17] "Online Violence against Women in Jordan, Realities and Recommendations" (in Arabic), Jordan Open-Source Association.

[18] https://women.jo/~women/sites/default/files/2022-06/اللجنة الوطنية لشؤون المرأة، دراسة العنف ضد النساء في المجاليين العام والسياسي.pdf
العنف ضد النساء في المجالين العام والسياسي، 2022.

[19] "Gender Based Violence Risk Assessment for East Amman", GBV Sub Working Group, October 2021.

3.1. Context

In April 2022, the Jordanian Council of Ministers approved the first child rights law in the Kingdom, and referred it to the House of Representatives, to go through the required legislative stages during an extraordinary session, held from 20 July to 29 September.

There were a number of compelling reasons for the government's submission of the draft law, as stated by the members of the Government themselves, including the Kingdom's fulfilment of its international obligations relating to the Convention on the Rights of the Child, which the Kingdom ratified on 24 May 1991,[20] and also in line with the constitutional amendments recently approved by the House of Representatives,[21] including the fifth paragraph of article 6 of the Jordanian Constitution, on the prevention of the abuse and exploitation of mothers, children, and the elderly.[22]

The draft law included provisions related to the protection and care of children in key sectors, such as education, health care, alimony, and custody. While these had already been included in other laws, such as the Personal Status and Education Laws, this draft included new provisions related to recreation, protection from forced labor, begging, addiction, and the provision of legal assistance to children.[23]

The debate on the draft law took place beginning with the opening session of the 19th extraordinary session of the Assembly, on 20 July 2022, and lasted two months, before the law was approved by the National Assembly, both deputies and senators, on 27 September.[24] During this period, the draft was subject to wide debate in parliament, as well as in society in general, becoming a prevalent issue on social media platforms, including in many targeted campaigns and through the circulation of disinformation and hate speech. These were based on demands for the repeal of the law as a whole and questioning the reasons behind it, as well as demands by the majority of deputies for careful study of the articles of the draft, by referring it to the competent committees to contribute to its improvement and better define of its terminology. Others suggested that representatives of the General Iftaa Department (a government institution) [25] attend sessions on the draft to ensure that its provisions were compatible with Islamic law and the traditions of Jordanian society. This ultimately

[20] "Press release: The Government Approves the draft Act on the Rights of the Child" (in Arabic), website of the Ministry of Social Development.

[21] Omar Hamza, "Research Article: Jordan's Role in Implementing a Commitment Declaration on Childhood Issues in the OIC Member States" (in Arabic), Ministry of Social Development website.

[22] "The Jordanian Constitution and its Amendments for the Year 2022" (in Arabic).

[23] Omar Ajlouni, "Research Article: "Child Rights Law in Jordan. A Positive Step But?" (in Arabic), Euro-Mediterranean Human Rights Monitor.

[24] "Press release: Senate Approves Child Rights Law" (in Arabic), Senate of Jordan website.

[25] The Jordanian General Iftaa Department is the body responsible for issuing fatwas according to the Iftaa Law No. 60 of 2006 and its amendments: Statement of the legal ruling in any matter of public and private affairs, the law is available at: https://www.aliftaa.jo/Default.aspx

happened, as the draft was referred to a joint committee (law and family) in the presence of Sharia judges from the General Iftaa Department.[26]

This report presents an analysis of the discourse on the Children's Rights Act, which has been characterized by much confusion and many accusations and divisions, as well as hate speech and misleading information, on a number of public pages on Facebook. The analysis was performed according to a specific methodology and monitoring carried out from 20 August to 10 October 2022 to provide a rich theoretical context for the case. It also allowed for the formulation of a group of recommendations for actors at the national and regional levels, to contribute to transparent regulations that address online hate speech and disinformation.

The Act, since it was introduced as a bill, was surrounded by great controversy, especially as it followed a previous controversy that revolved around the consequences of implementing the Islamic Centers Law No. (107) 2022,[27] as this was the starting point for some opponents of the draft Children's Rights Act. For these critics, the draft Children's Rights Act reinforced the prevailing negative impression about the introduction by the state of measures, imposed by international organisations and external bodies, that contradict the system of values and customs in society.[28]

On social media platforms, there were two main groups active in discussions of the draft, the first in opposition to the measure, including conservative and Islamic parties, social movements, and other groups and public figures, and the second in support, including civil and left-wing parties and movements, human rights activists and former state officials.

3.2. Methodology

The research team analysed the content of 10,188 comments across specific Facebook pages. The share of those classified as uncontroversial was 66.43, while 2,085 comments contained hate speech, equivalent to 20.47 per cent of all comments analysed, another 10.6 per cent of comments were identified as containing misinformation, and the share of those classified as falling into "other categories", i.e., those containing advertising or comments unrelated to the post in question, was 2.5 per cent.

When classifying hate speech comments, it was found that denigration was the

[26] Press release, Chairman of a Joint Parliamentary Committee: "Studying the Articles of the Draft Act on the Rights of the Child Thoroughly" (in Arabic), the website of the Kingdom Channel.

[27] As a result of the implementation of the Islamic Centers Law No. (107) 2022, the Ministry of Awqaf and Islamic Affairs and Holy Places suspended the work of 30 Quranic centers, affiliated with the Society for the Preservation of the Holy Quran, the case witnessed the circulation of accusations between the two parties that extended to social media platforms and included different directions, in June 2022:. "Press release, Awqaf: We Have Partially and Temporarily Suspended 30 Violating Islamic Centers", Petra Agency Jordan; "Definition in the Association" (in Arabic), website of the Society for the Preservation of the Holy Quran; "Video clip, Voice of the Kingdom| Associations for the Preservation of the Qur'an and Preparation for the Hajj Season, the Voice of the Kingdom Program" (in Arabic), YouTube: "Kingdom Channel", "Video clip, Voice of the Kingdom| Associations for the Preservation of the Qur'an and Preparation for the Hajj Season, the Voice of the Kingdom Program" (in Arabic), YouTube: "Kingdom Channel".

[28] MP Yanal Freihat, writer Tarek Delawani, explaining the reasons for opposing the law, available at: https://www.facebook.com/yanal.fraihat1/videos/845804559719148 / https://b.link/gkcghm

most common form, accounting for 7.54 per cent of all comments characterised as hate speech, followed by defamation (6.16 per cent), insults (4.20 per cent), and cyberbullying (1.2 per cent). The table below provides a full breakdown of the classifications of hate speech, their number and share as a percentage of all comments monitored:

NUMBER OF REVIEWS	PERCENTAGE OF TOTAL COMMENTS CONTAINING HATE SPEECH	CLASSIFICATION
20	0.20%	Violence
628	6.16%	Defamation
768	7.54%	Denigration
428	4.20%	Swearing
123	1.21%	Cyberbullying
20	0.20%	Sexual Harassment
65	0.64%	Incitement
33	0.32%	Discrimination
2085	20.47%	Total

Figure 25: Categorization of hate speech comments on the children rights bill

RATE	INTENSITY CLASSIFICATION	DESCRIPTION
0	No Intensity	Content that is not related to the post and does not incite in any way.
1	Disagreements	An expression that reflects a difference of opinion in relation to an idea, belief, etc.
2	Negative Actions	An expression containing non-violent actions associated with a group or party, or responses containing non-violent actions, such as metaphors. Examples include accusations of theft, threats, indecency, mistreatment and alienation.
3	Building a negative	An expression containing a non-character violent characterisation and insults, such as accusations of stupidity, robbery, counterfeiting, insanity,.
4	Demonizing and	An expression containing inhumane Dehumanising and characterisation of inferiority, such as the use of labelling words associated with animals, diseases and others.
5	Violence	An expression that involves inflicting physical or metaphorical harm, inciting such harm, and responses that call for physical or metaphorical violence, such as torture, rape, beatings, etc.
6	Death	An expression that includes the word “murder” by a particular group, and responses that involve murder.

Figure 26: Hate speech intensity scale, developed by Al Hayat to classify hate speech content.

3.3. Data Analysis

3.3.1. Tools for directing public opinion

During the research, several entities and people were monitored who directed public opinion on the draft Children’s Rights Act, whether in support or opposition. This was done using several tools, including technical tools and social media employing religious and/or ideological discourse, as well as discourse based on national traditions. Some of these utilised multiple tools, while some were satisfied with just one tool. The different tools are covered below.

Digital tools

Public figures have used various digital tools to guide public opinion in opposition to or support of the draft Act:

A. Creating websites, such as one called "The Child Law is Poisoned", which was one of the tools used in the campaign by opponents of the law. The site provided information on its campaign, as well as photos, videos and documents from its work, and allowed viewers to sign up as volunteers within the campaign.[29]

B. Creating official campaign accounts on social media platforms, where a page for "The Child Law is Poisoned" was created on the Facebook, YouTube and Twitter platforms.[30]

C. Creating channels or playlists on the YouTube platform, including an operational list created by Dr. Iyad Al-Qunaibi on his official "YouTube" channel, entitled "The Child Law - War on Nature".[31]

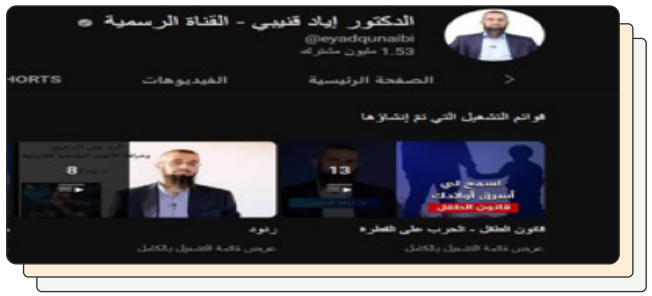


Figure 27: A screenshot of the Facebook page of Dr. Iyad Qunaibi on Facebook, with 1.53 million followers

D. The use of public personal accounts on social networking sites by a number of public personalities.[32]

E. Calling for "cyber storms" on through social networking sites, which emerged on the Twitter platform.[33]

F. Using live broadcasts and hosting speakers, particularly on Facebook.[34]

G. The use of hashtags, including "For the child law", "Poisoned Child Law", "approving the law is a shame", and "Jordan against the child law".

[29] The website was suspended, and this was its link: https://www.childlaw.info/

[30] Dr. Iyad Qunaibi's Facebook page is available at: https://www.facebook.com/childlawjo/?ref=page_internal

[31] Dr. Iyad Qunaibi's YouTube page is available at: https://www.youtube.com/@eyadqunaibi/playlists

[32] The account of Dr. Mohamed Tohme Al-Qudah on "Facebook", available at: https://www.facebook.com/profile.php?id=100048780401429

[33] Call for a Cyberstorm, available at: https://b.link/27mr3w

[34] Live broadcast on the issue of the bill of law on the rights of the child, available at: https://b.link/x767ru

Intellectual tools

A- Employing religious discourse

Some actors have worked to spread their ideas and influence public opinion on the draft Act with religious discourse. Preacher and academic Iyad Al-Qunaibi and some Islamic activists have used this type of tool in an attempt to win public opinion to exert pressure to rejecting and dropping the draft. They used various digital tools, and sharp speech that stimulates religious feelings. In some cases, this went beyond conveying opposition to the draft bill and pushed for the circulation and dissemination of hate speech and unverified information, whether through comments on campaign publications or by circulating and advocating for certain content.

The most prominent categories of this speech are defamation, denigration, cyberbullying, insults and slurs.

When looking at the comments that were monitored, the work team found a repetition of the relevant words in religious discourse, for example, the word "religion" was mentioned in 873 comments, the word "Islam" 547 times, and the word "Qunaibi" 86 times, out of 10188 comments that were monitored. This reference to religion may have been a tactic to attract more followers. An example of this type of speech, which was published by several public figures, reads: "Everyone before today and silent today is an accomplice in the crime and responsible before God for it." Another, shown in the image below, and this use is evident from the title of the video clip, which is part of the campaign that al-Qunaibi worked on.[35]

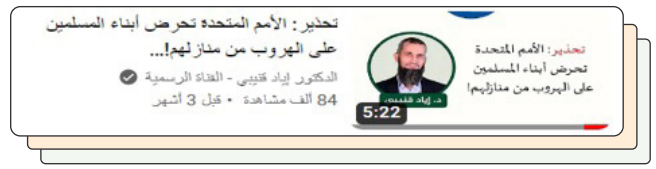


Figure 28: Screenshots from a video by al-Qunaibi, with the title "Warning: United Nations is encouraging children to escape from their homes"

B- Employing religious discourse in an extreme manner

One of the most prominent cases monitored in the research was the speech of former MP Muhammad Tohme Al-Qudah (of the Islamic Movement), who said in a live broadcast on his Facebook page that "The bill of law on children rights includes articles that those who approve and approve them will have come out of the religion and disbelieve, because they contradict Islamic law, which are Articles 8, 20 and 21."[36]

[35] A video clip on Dr. Iyad Qininbi's page, available at: https://b.link/2wqcjd

[36] A video clip from Dr. Mohamed Tohme Al-Qudah, available at: https://b.link/o3wm3a

Figure 29: Screenshot from the live video published on the Facebook page of former MP Muhammad Tohme Al-Qudah (of the Islamic Movement)

The working group identified a repetition of words related to extremist religious discourse in the monitored comments, for example, the term "infidel/infidelity" appeared in 67 comments, and the term "CEDAW [37]" in 252 comments, out of a total of 10,188 comments observed.

Employing the national security discourse

The analysis revealed that some figures who previously held roles in the state apparatus engaged in a national security discourse against opponents of the law (especially those from the Islamic Movement). Most notable among these was a member of the Senate and former MP, Jamil Al-Nimri (Secretary-General of the Social Democratic Party), who attacked what he called "this obscurantist vision and its supporters", suggesting the need to confront those opposing the draft Act. He also wrote in a piece in the Al-Dustour newspaper that "[t]his storm that claims against the law what is not in it or carries texts of implicit meanings that are intolerable will be revealed tomorrow in the responsible legislative debate in the committees of the Council, but the media campaign wants to preempt this by demonising the law and setting it a target for bombing with all kinds of weapons, which seems to be a kind of rough messages to the state and feel the pulse of entering into an arm-twisting battle with countries, and we thought that we have exceeded the time of these methods." [38]

Former MP Qais Ziadin, from the Civil Movement party, wrote in an article in the official al-Ghad newspaper that "... [p]ublic opinion was misled and mobilised not only against the law, but against the state. The emergence of a harsh treasonous discourse against the 'state' is very dangerous, especially since apolitical citizens are convinced, and today the danger has become an attempt to eliminate trust or belonging between the citizen and his state and not his[39] government, and the difference is great."

When looking at the 10,188 comments monitored, the working group found a repetition of related words in the national security discourse, for example, the terms "homeland" and "country" appeared in 275 comments, while the terms "darkness", "ISIS" and "terrorism" appeared in in 21 comments.

3.3.2. Trends in analysing the interaction with the issue of the Children's Rights Act on Facebook

- The issue of closing Quran memorisation centres and the Child Rights Law was linked to the fact that the government's actions in both cases was viewed by opponents as a war on religion and the traditions of the Jordanian people.
- CEDAW is linked to the law by its opponents, stating that its passage and approval will serve the CEDAW agenda, and claiming that it will be the beginning of the introduction of laws that support and legalise homosexuality in Jordanian society and will lead Jordanian society towards Freemasonry and secularism.
- Promoting the content disseminated by the preacher Iyad Al-Qunaibi, by publishing links to his videos that explain his belief the draft Children's Rights Act is dangerous and that it is against human nature. It has been observed that some of the content that interacts with the discussions on the bill comes from outside Jordan, particularly from the Gaza Strip, the Arab Republic of Egypt and Sudan. These interactions mostly align with viewpoints opposing the bill and frequently involve the use of religious language.

Examples from users sharing the links to Al-Qunaibi's channel on YouTube:

- "Dr. Iyad Al-Qunaibi excelled by explaining the catastrophic child law on his YouTube channel, #Child_ Law_is_poisoned episodes published by Dr. Iyad Al-Qunaibi so far on the subject: 1. Allow me to steal your children: https://youtu.be/S8AUfOji-OI
- Responding to the defenders of the Child Law: https://youtu.be/zA9p2HihqpA

"Dr. Iyad Al-Qunaibi's response to the law of Faltan and the destruction of the child https://youtu.be/S8AUfOji-OI ; https://youtu.be/Tl7l2R2xFUQ https://youtu.be/zA9p2HihqpA ."

"Al Jazeera has published a report entitled: Jordan – Controversy over the draft child law which is available at a link that we will put below.[40]

[37] For some religious extremists, it is a text that leads to infidelity, some religious extremists, is seen as an incitement to infidelity.

[38] Jamil Al-Nimri, "When a Child is a Victim of Politics" (in Arabic).

[39] Qais Ziyadin, " Where are the Statesmen?!" (in Arabic).

[40] Link to Al Jazeera's report, available at: https://b.link/75pmes

The report shows either extreme superficiality or unprofessionalism! So, it presents the reasons for opposing the law in a poor way that never shows the magnitude of the danger that we and others have repeatedly demonstrated. We ask those who have watched the episodes that we have published on the subject so far, or some of them, to comment on the report on the Al Jazeera Live page by indicating the episodes and the content of what is in them, and how the phrases set in the law are an entry point to systematically corrupt our children, with the use of the hashtag #Child_LawPoisoned.

3.3.3. Analysis of hate speech related to the Children's Rights Act from a gender perspective

In research and monitoring, it is necessary to address gender issues and the related societal roles, and what the context is in which gender data is addressed. The monitoring team of Al Hayat Rased looked at the gender perspective at different levels, particularly whether gender was the main topic or a secondary factor.

1. The social role associated with men in the family in Jordan, where many comments on the issue of the Children Rights Act emerged and were rejected by men because of the roles required of them to manage family affairs, and that this law threatens their ability to maintain the cohesion of the family, poses a threat to family stability and their roles, specifically, as men.
2. Creating a link between women's and children's rights at the expense of men, presenting the bill as another step towards family disintegration and creating space for women and children to bypass men and threaten their status. The law was seen as a complementary step to previous endeavours to provide rights for women, thus reducing men's control over their families, an initiative that was widely rejected by some online actors.

Al Hayat Rased studied several examples of hate speech campaigns against women politicians in Jordan. There were similar trends in how the retrograde and sexist speech is rooted in religious and social backgrounds to attack former MPs and human rights defenders for their position in favour of the child law.

Below are some examples (figure 30 to 34) retrieved from the monitoring of comments found in the accounts of:

Mrs. Roula Al Hroub: A former member of the 17th parliament from the capital, Amman, and a professor at the Faculty of Educational Sciences at the University of Jordan, known for her opposition to successive governments in the Kingdom.

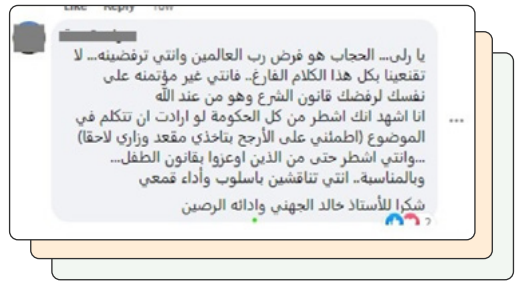


Figure 30: Screenshot of a comment on Al Hroub post criticising her for not wearing a hijab, and also questions her capacity to give her opinion on religious rules.

Figure 31: Screenshot of a comment attacking Al Hroub for her non-Jordanian origin.

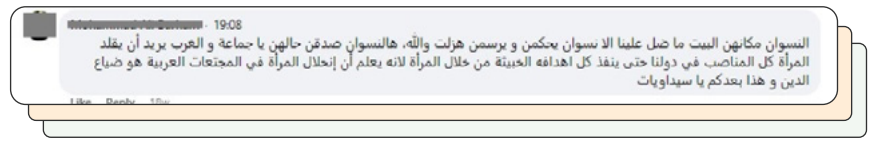


Figure 32: Screenshot of a content attacking woman's rights advocates.

In this second example, the monitoring team of Al Hayat retrieved comments on Ms. Hala Ahed page. Ms. Ahed is a lawyer and human rights activist in the field of defending women and prisoners of conscience. She was part of the legal team defending the Jordanian Teachers' Syndicate in its dispute with the government in 2020. She also chaired the Legal Committee of the Women's Union in Jordan.

Figure 33: Screenshots of two comments attacking Ms Ahed for sharing thoughts that "are not suitable for a woman who wears Hijab" and that "women should never rule"

Finally in this third example, Dr. Nofan Al-Ajarma, a former minister and former head of the Legislation and Opinion Bureau (a government institution), some comments were also attacking him for "spreading homosexuality, atheism and pornography"

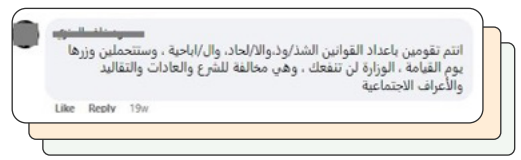


Figure 34: screenshot attacking Dr. Nofan Al Ajarm

Conclusions

Al Hayat-Rased's analysis found that the extent to which Jordanian users respond to rumours and inaccurate information through social media platforms in the accounts of activists and influencers is greater than their response to official data and responses issued by official figures and institutions. This matches the findings of an earlier opinion poll conducted by Al-Hayat-Rased- entitled "Measuring the Level of Jordanians' Knowledge of Hate Speech and Misleading and False information on Social Media Platforms."[41] The results of this earlier opinion poll showed that :

- Seventeen per cent of Jordanians said that they do not have time to verify the truth of the news they read, 13 per cent said that they do not know how to verify the authenticity of published news, 18 per cent said that they share news published by people or entities they trust, while 17 per cent said that they publish news from specialised bodies.
- There is a crisis of confidence suffered by official institutions to convince their discourse and narrative at the expense of the wide impact of the speeches of activists and influencers, and this gap is increasing.
- Jordanians showed a high response to conspiracy theories, including that the religion and cohesion of the family and society are being targeted by Freemasonry, Zionism, and unknown foreign parties.

The following terms were widely used to spark emotional religious discourse and stir up religious feeling by opponents of the law and, thus, to influence public opinion: "dismantling the family", "atheism of the child", "changing his religion", "sexual frenzy", "homosexuality", "taking away the identity of the Islamic community", "taking the child from his family," and "Western domination of our families and traditions".

The most widely circulated forms of hate speech, bullying, insults, defamation, and incitement to violence have been used against supporters of the Children's Rights Act, activists, and former officials.

A shift in political orientation from centrist to conservative was observed among some members of the current House of Representatives and former officials when commenting on this issue.

There is a lack of trust in and questioning of the work and motives of international organisations, as well as civil society institutions cooperating with them, viewing them as challenging the culture and cohesion of society.

The results showed that 20.47 per cent of the total comments analyzed (10,188) contained a form of hate speech.

[41] Al Hayat- Rased releaseda Study on "Measuring Jordanians' Level of Knowledge of Hate Speech and Misleading and False Information on Social Media Platforms", available in Arabic here: https://cutt.ly/PNpv88G

It was found that denigration were the most common form of hate speech, at 7.54 per cent, followed by defamation, at 6.16 per cent, insults, at 4 Per cent, and cyberbullying, at 1.21 per cent.

Some commenting used national security discourse to counter opposition to the draft Children's Rights Act, and by using defamatory skeptical rhetoric.

The role associated with men in the family in Jordan was raised regularly, where many comments by men rejected the Children's Rights Act, saying they are naturally burdened by the obligations on their shoulders to manage the affairs of their family, and that this law only threatens their ability to maintain the cohesion of their family, and poses a threat to family stability and their roles specifically as men.

Creating a negative link between the rights of women and children at the expense of men was put forward as another issue that could lead to family disintegration and making space for women and children to bypass men and threaten their status. Some comments considered the law as a complementary step to previous measures giving rights to women, and thus contributing to men losing control over their families, and this was widely rejected by many. of those commenting

Recommendations

- Provide a recommendation to ministries and official departments to increase interaction with the public through social media platforms, to deliver updates and explanations about legislation to citizens in a simple, clear, and accessible manner, in addition to using social media as approved platforms to provide statements and news of interest to citizens, allowing these to be their reference, rather than sources that contain misleading information.
- Start consultations with the concerned authorities – the Ministry of Justice, the Ministry of Digital Economy, the Legislation and Opinion Bureau and the Cybercrime Unit – to reach and adopt a clear and comprehensive definition of hate speech in Jordanian laws, so that it is consistent with international standards and preserves freedom of expression.
- Submit a recommendation to the Ministry of Digital Economy to develop the level of relationship between it and Meta Platforms (formerly Facebook) through the latter's office in Jordan, so that the Ministry can participate in developing the company's social media platform policies in line with Jordanian legislation and best practices to address hate speech and disinformation, using the work of CSOs working on this issue.
- Direct CSOs to organise a series of national workshops on legislative amendments needed to better address the challenge of hate speech, disinformation, and false news. These should include the participation of judges dealing with these legal issues.

New emerging threats in Disinformation

Interview with Lena-Maria Böswald.

Interviewed by Wafaa Heikal

Hey Lena, it's a pleasure to be in conversation with you today. The Words Matter network is keen to connect with local and international researchers about disinformation and hate speech.

Can you please tell us a little bit about your background and your current role with DRI?

Hi Wafaa, thanks for having me! I am a Digital Democracy researcher at DRI Berlin, conducting research on new emerging threats in the disinformation field, derogatory speech and hateful content. Focusing on democratic processes and election monitoring online, I also help build social media monitoring capacity for EU election observers. Before I joined DRI, I worked in Communication Science at the University of Amsterdam. I also touched on disinformation early on in my early research but, back then, it was still called "fake news".

We have been following your work on new emerging disinformation threats – can you describe the main trending threats?

In our approach at DRI we focus on three different dimensions in the sphere of disinformation. We focus on the technical foundation of disinformation content – the tools that distribute disinformation (fake images, deepfakes, synthetic audio); then on tactics – strategies used to propagate disinformation content (domestic proxies, shadow websites, cross-platform sharing); and then on narratives – the stories that combine the tools and the tactics to distort facts (gendered disinformation using cheapfakes and deepfakes).

We have seen that in recent disinformation campaigns innovations in tactics have played a much bigger role than innovation in the tools themselves. At DRI, we focus a lot on synthetic content produced with the help of AI as a disinformation tool, but in many elections around the world we have seen it play only a marginal role so far. It is still far more common to find less computationally advanced cheapfakes than sophisticated deepfakes, due to lack of expertise and resources.

If malicious actors can use fewer automatic techniques that produce effective content and can go viral, there is less incentive for them to invest in creating more complex synthetic media. We have, however, seen some rapid developments and advances in the last couple of months in the creation of synthetic media – easier access to AI-generated content, more refined models and the combination of multiple tools. There have been rapid advances in what we call “generative AI”, models that seem to display humanlike activity, such as text-to-image generation (e.g., DALL-E2 and Stable Diffusion), but also text generation of large language models that can predict language and can produce human-like text based on a simple prompt that you give the machine learning model (e.g., ChatGPT).

The merging of these tools is something to look out for. Now that they can be combined, that makes disinformation campaigns and fake evidence for false content much more believable and more difficult to debunk. But we have also seen new tactics emerging across the board in the last couple of months, such as strategically targeting institutions of trust, be it the media or governmental bodies that used to be trusted, using mirror or shadow websites to imitate news outlets, or laundering information through proxies or junk websites, so not actually producing the content themselves, but making use of other people's resources.

These are tactics and tools that are already applied and that, with easier access, will become more prominent in the future, so that's something to prepare for.

What should MENA researchers be vigilant about and prepared for when it comes to new, emerging threats?

Be on the lookout for copycat patterns in the MENA region; there is the tendency to look carefully for strategies in other countries and adapt similar strategies or narrative patterns that have worked before in a different context. This could be the efficient use of cheapfakes, disinformation disguised as satire, of face-swapping apps, for example. It is often the case that authoritarian leaders learn from each other; if a tactic or technique works in one country, it can be easily applied to another country and fed into existing political narratives.

How can AI-powered disinformation campaigns impact democracies around the world?

That's a very tough question because it's difficult to predict. AI-powered disinformation campaigns have the potential to deepen the harm that is already posed by disinformation campaigns. They can not only reshape the speed and quantity, but also the quality of spreading disinformation. It's the sheer volume of disinformation that can be easily produced and easily shared with the help of AI. There's a very strong likelihood with evolving technology that this gap between fake but relatively plausible content and authentic content based on facts can be narrowed and, therefore, can influence democratic discourse.

What happens if everything can be fake? Nothing must be real – and this will be a big issue for democracy, because when people think that everything they see can be fake, how can they figure out what's real?

We are entering a phase, with all these large language models, where false content can also be entirely synthetic, thus allowing disinformation actors to use synthetic text as a basis or a foundation to create fake imagery or false evidence. It's this automatisation and the very synthetic nature of disinformation production that replaces the very tedious act of creating content from scratch and increases the complexity of disinformation campaigns.

That's what is most worrisome – how AI affects democratic discourse. Everything is getting more complex, and it's more difficult to decipher what is wrong and what is right.

At DRI, how are we building our capacity and sharpening our tools to counter these new emerging threats?

Our main pillars are Foresight, Raising Awareness and Pre-Bunking.

At DRI, we are exploring emerging challenges to online political speech. We try to warn by using early detection mechanism and a lot of our work focuses on pre-bunking – preparing society for specific disinformation efforts that have or may not have happened yet, making sure people understand how a certain technological advancement could be used for good, but can also be used for bad in disinformation campaigns. It is just making people aware of the malicious face of the tool.

And what are the main recommendations that CSOs working in information integrity and strengthening democracies should focus on?

First, raising awareness, so that people can be watchful for manipulated content, and also educating them about the prospect of new tech being used in disinformation campaigns.

Second, effective collaboration between researchers, CSOs, and, most importantly ,tech companies. It is difficult to get them on board, but you can only inform the public debate if you have a very strong ecosystem and follow developments in the field to build resilience around civil society. This is only possible through collaboration and making sure that different stakeholders are all on the same page.

How can disinformation (in general) be gendered? What will be the impact of AI on this?

On a global scale, we have seen the use of disinformation campaigns targeting women on a regular basis, and not just in politics. There is a tendency to disseminate gendered disinformation, spreading deceptive or incorrect information targeting women in society by, for example, questioning their competence and their fitness for office, if it's a political figure. There is a lot of online gender-based sexual violence, and the sole aim is to discredit and silence women, to ensure that women simply no longer participate in political discourse.

AI can move gendered disinformation to the next level because AI already works with existing biases. AI is often trained on data that's inherently biased and reproduces gender stereotypes; that's one issue. With the increase of AI, we also see that deepfakes are easy and cheap to produce and can be used to humiliate women, to discredit them, often in the form of using non-consensual pornographic content as a means of disinformation. With the advance of new technology these deepfakes are going to be more professional, the quality will improve, and then it will be extremely hard for victims to prove whether it's them or not in a video or an image.

This is moving gendered disinformation to a completely different level, and it leaves the victim alone, as they have the burden to chase after the content, as the content can't always be removed in its entirety, because other users upload it again.

I am not quite sure whether we are prepared for it. We have the detection mechanisms, authentication infrastructure and forensic tools to detect a lot of manipulated content, but it takes a lot of time, effort, self-regulation by service providers and, often, still human content moderation to do so. The damage that is being done to victims is unimaginable.

Platforms need to get better at stopping the harm before it's widely circulated, by limiting the algorithmic exposure to content (e.g., slowing down the circulation of content before removing or deranking it). AI service providers should include authentication infrastructure at the point of creation of images and videos (e.g., provenance technology) and regulators need to try to find more ways to mitigate the production of such malicious content (e.g., stronger regulation of deepfake apps that can be used to produce non-consensual pornographic content) in the first place.

My last question for you: we found a few women working in our domain in countering disinformation and hate speech.; how can we get more women researchers in the global majority especially in MENA?

It's always easier said than done. I think the crucial aspect is to provide women with the technical skills to analyse discourse online to make sure that they have a seat at the table. If the skills aren't there, it is harder to get represented. So technical skills are like a key civil society can push and ask for that CSOs can provide. That's not an easy task and not a simple solution but, for me, it is still the only effective one to provide people with technical skills in the first place.

Bibliography

Reports and studies

European Union, Election Observation Mission To Lebanon 2022, May 2022. Available online here: ممارسات شراء الاصوات اثرت في حرية اختيار الناخبيين وادت الى غياب في تكافؤ الفرص. | EEAS Website (europa.eu)

UNDP, 2022 Lebanese Parliamentary Elections: Key Results, December 2022. Available online here: 2022 Lebanese Parliamentary Elections: Key Results | United Nations Development Programme (undp.org)

Democracy Reporting International, Online Public Discourse in MENA: Disinformation and hate speech during the 2022 Lebanese and Jordanian elections, 28 September 2022. Available here: https://democracy-reporting.org/en/office/global/publications/word-matters-published-the-first-report-on-disinformation-and-hate-speech-in-mena-regions

Maharat Foundation, Political propaganda and information manipulation on social media during the Lebanese Parliamentary elections 2022. Available here: DRI REPORT ENGLISH.docx (maharatfoundation.org)

Jordan Open-Source Association – Online violence against women in Jordan, realities and recommendations, January 2022, Available in Arabic here: https://josa-api.fra1.digitaloceanspaces.com/portal/c952aa7cc31444a5fa07be4ab385d870.pdf 2021 ،مقترحة (ورقة سياسات).

The Jordanian National Commission for Women, violence against women in public and political sphere, June 2022. Available in Arabic here: https://women.jo/~women/sites/default/files/2022-06/دراسة العنف ضد النساء في المجاليين العام والسياسي.pdf، اللجنة الوطنية لشؤون المرأة، العنف ضد النساء في المجالين العام والسياسي، 2022.

Al Hayat Rased, Research study on knowledge of hate speech and disinformation on social media by Jordanians, August 2022. Available in Arabic here: راصد يصدر دراسة حول «قياس مستوى معرفة الأردنيين/ات بخطاب الكراهية والمعلومات المضلّلة والزائفة على منصات التواصل الاجتماعي» | راصد (rasedjo.com)

Articles

Omar Ajlouni, Research Article, Child Rights Law in Jordan. A positive step but? Euro-Mediterranean Human Rights Monitor, available at: https://cutt.ly/7BByRu0

Omar Hamza, Research Article, Jordan's Role in Implementing a Commitment Declaration on Childhood Issues in the OIC Member States, Ministry of Social Development website, available at: http://www.mosd.gov.jo/ui/arabic/ShowContent.aspx?ContentId=487

Press release, Senate approves Child Rights Law, Senate website, available at: https://cutt.ly/aBBuRK

The Jordanian Constitution and its amendments for the year 2022, available at: https://b.link/x3fyb1

"Introducing New Ways to Create, Share and Discover Live Video on Facebook", in META news, 06 April 2016. Available here : https://about.fb.com/news/2016/04/introducing-new-ways-to-create-share-and-discover-live-video-on-facebook/

Press release, the government approves the bill of law on the rights of the child, the website of the Ministry of Social Development, August 2022. Aavailable in Arabic at: http://mosd.gov.jo/UI/Arabic/newsview.aspx?NewsId=2475

Media and websites

- Al Jazeera Jordan
- Al Ghad
- Al Nahda News
- Digital Monitor
- Nidaal Watan
- Addustour
- Jordanian Senate
- Jordanian Ministry of social development

About Words Matter

Contact: wordsmatter@democracy-reporting.org

DRI has been increasingly active in the field of social media monitoring (SMM) since 2017, strengthening local capacities to monitor social media during elections, sharing information and evidence gathered in different countries, bringing together expert organisations, producing methodologies, and informing public and expert debate.

Within the framework of the "Words Matter" project, DRI and its partners seek to contribute to strengthening the safeguarding of democratic processes and societies' resilience to online disinformation and hate speech in the MENA region.

DRI works with partner organisations from four countries (Jordan, Lebanon, Sudan, and Tunisia), strengthening local capacities to monitor and analyse online disinformation and hate speech during key national democratic processes, while building a regional network to allow for comparative analysis and peer learning.

"Words Matter" aims to achieve the following objectives:

- Capacity-building for project partners **to acquire institutional skills to design sound social media monitoring methodologies**, to effectively monitor disinformation and hate speech online, and to enhance evidence of the impacts of disinformation and hate speech online on civic or political participation and human rights.
- **Enhanced multi-stakeholder and regional engagement** to advocate against and combat online disinformation and hate speech, through a civil society network, as well as through continuous exchanges on transparent regulations.; and
- In the countries of project partners, **improved awareness and resilience of civic target groups**, and concrete action by decision-makers to transparently combat online hate speech and disinformation.

About the Digital Democracy Program

Contact: info@democracy-reporting.org

DRI's Digital Democracy (DD) programme protects online democratic discourse by exposing disinformation, manipulation and hate speech, strengthening the capacity of CSOs for monitoring and advocacy, and ensuring appropriate and evidence-based responses from governments and tech companies.

DRI is well-positioned to address online threats and disinformation, due to its research on manipulated media content, deepfakes as potential disinformation tools, and its current focus on identifying new potential threats and emerging technologies in this field. As part of our diverse toolbox, we have, for example, integrated machine learning models to help us identify emerging trends in the disinformation space. Our work on information manipulation is also complemented by analysing and publishing guides on gender-based under-representation and harassment online.

An important activity within the DD programme for exposing and fighting hate speech and disinformation is social media monitoring (SMM). SMM is the objective analysis of democratic discourse and political actors on social media platforms. This is far more complex than traditional media monitoring, with a myriad of actors and content, combining official democratic institutions (e.g., political parties, politicians, media) and unofficial actors (e.g., individuals, political influencers, partisan groups). This is why DRI published the Digital Democracy Monitor Toolkit, the first social media monitoring methodology that helps civil society, journalists, and academia to research social media and democracy.

Our methodology was tested and used for conducting social media monitoring in 12 countries (including Germany, Libya, Myanmar, Nigeria and Sri Lanka), focusing on disinformation, hate speech and political advertising before, during and after the elections. By using a holistic approach to analyse social media, our toolkit engages with disinformation and hate speech by looking at the message or content, the active messengers, and the messaging, thus both the forms and the channels of distribution.

Based on the findings of our SMM, we have advocated for the implementation of the European Democracy Action Plan (EDAP) commitments, which could strengthen the fight against disinformation at the EU level and contribute to the debate about content-ranking systems, a major challenge when it comes to the dissemination of dis/misinformation. DRI has also lobbied for the implementation of the EU's Digital Service Act, a potential milestone in the effort to increase accountability across social media platforms. In launching the Arabic version of the SMM toolkit, we hope to empower the MENA region in the same way.

About DRI

Contact: info@democracy-reporting.org

Democracy Reporting International (DRI) is an independent organisation dedicated to promoting democracy worldwide. We believe that people are active participants in public life, not subjects of their governments. We strengthen democracy by supporting the institutions and processes that make it sustainable, and work with all stakeholders towards ensuring that citizens play a role in shaping their country. Our vision is grounded in globally agreed upon principles of democracy, stemming from the democratic governance championed by the United Nations and international law.

DRI's work focuses on five key themes of democracy: Justice, Elections, Local Governance, Digital Democracy and Human Rights. By working at both the national and local level, we use five intervention approaches in our projects: awareness-raising, capacity-building, fostering engagement between different stakeholders, supporting the building of democratic institutions, and advising on the drafting and implementing of policies and laws.

DRI's work is led by a Berlin-based executive team and supervised by an independent board of proven democracy champions. DRI maintains country offices in Lebanon, Libya, Tunisia, Pakistan, Myanmar, Sri Lanka and Ukraine. Through our networks of country offices and partners, we are in a unique position to track, document, and report developments and help make tangible improvements on the ground.

About DRI Partners

Al-Hayat Center for Civil Society Development: is a non-governmental civil society organization founded in 2006. The center has expanded to become one of the leading NGOs in Jordan. Al-Hayat's overall mission is to promote accountability, governance, public participation, and tolerance in Jordan and the region within the framework of democracy, human rights, the rule of law, and gender mainstreaming in public policy and actions.

Jordan Open Source Association (JOSA): is a non-profit organization based in Amman, Jordan. The association is among the few non-profits registered under the Jordan Ministry of Digital Economy and Entrepreneurship. JOSA's mission is to promote openness in technology and to defend the rights of technology users in Jordan. JOSA believes that information that is non-personal – whether it's software code, hardware design blueprints, data, network protocols and architecture, content – should be free for everyone to view, use, share, and modify. JOSA's belief also holds that information that is personal should be protected within legal and technological frameworks. Access to the modern Web should likewise remain open.

Lab'TRACK: is a laboratory for monitoring, analysis and reflection on political disinformation phenomenon on social networks, in particular the Facebook network. The laboratory is a collaboration between Mourakiboun and IPSI.

MOURAKIBOUN: Mourakiboun is a domestic electoral observation network that was launched in 2011 and is today a key player in this field with multiple national and international partners. Since 2014, Mourakiboun has been diversifying its actions by adding accountability of public services and support to the Tunisian decentralization process to its portfolio. Mourakiboun has a network of over 100 volunteers in all regions of Tunisia and excellent access to local structures and stakeholders. Mourakiboun has adopted an IT approach to its activities, thereby increasingly reaching Tunisian youth. During the 2014 and 2019 presidential elections, Mourakiboun conducted social media monitoring activities focused on the interactions of FB users with the speeches of candidates during electoral campaigns.

Institut de Presse et des Sciences de l'Information (IPSI): was established in 1967 and became a

non-departmental public institution enjoying financial autonomy and legal personality in 1973. The Institute is known as Tunisia's leading university for the education of journalists and media workers. IPSI's research in the field of information and communication sciences has been met with international acclaim. IPSI has a network of national (INLUCC, HAICA, UFP) and international partners (Deutsche Welle Akademie, UNESCO, UNDP, Article 19 among others). Through this cooperation, IPSI provides specialized training sessions and hosts experts and internationally renowned speakers to introduce students to innovative practices in the field of communication.

MAHARAT: a women-led, Beirut-based organization, working as a catalyst, defending and advancing the development of democratic societies governed by the values of freedom of expression and respect for human rights.

Maharat advances the societal and political conditions that enhance freedom of expression and access to information, both online and offline. Maharat engages and equips a progressive community in Lebanon and the MENA region with the skills and knowledge necessary to create change.

Sudanese Development Initiative (SUDIA): Founded in 1996 and established in Sudan since 2002, the Sudanese Development Initiative (SUDIA) is a non-governmental, non-profit organization working with a broad cross-section of actors and stakeholders developing programs and providing services focusing on peacebuilding and community development security, natural resource management and environmental conservation, democracy, and human rights promotion. SUDIA has experience monitoring both online and offline media. Since 2006, the organization has carried out diverse projects and activities that target the media sector ranging from media research and assessments, convening of discussion forums and roundtables, training of journalists, and extending support to journalists and media development initiatives in the country. More recently, as part of the Sudan Hate Speech Lexicon project being implemented jointly with Peace Tech lab, SUDIA has been monitoring the use of hate speech on social media."

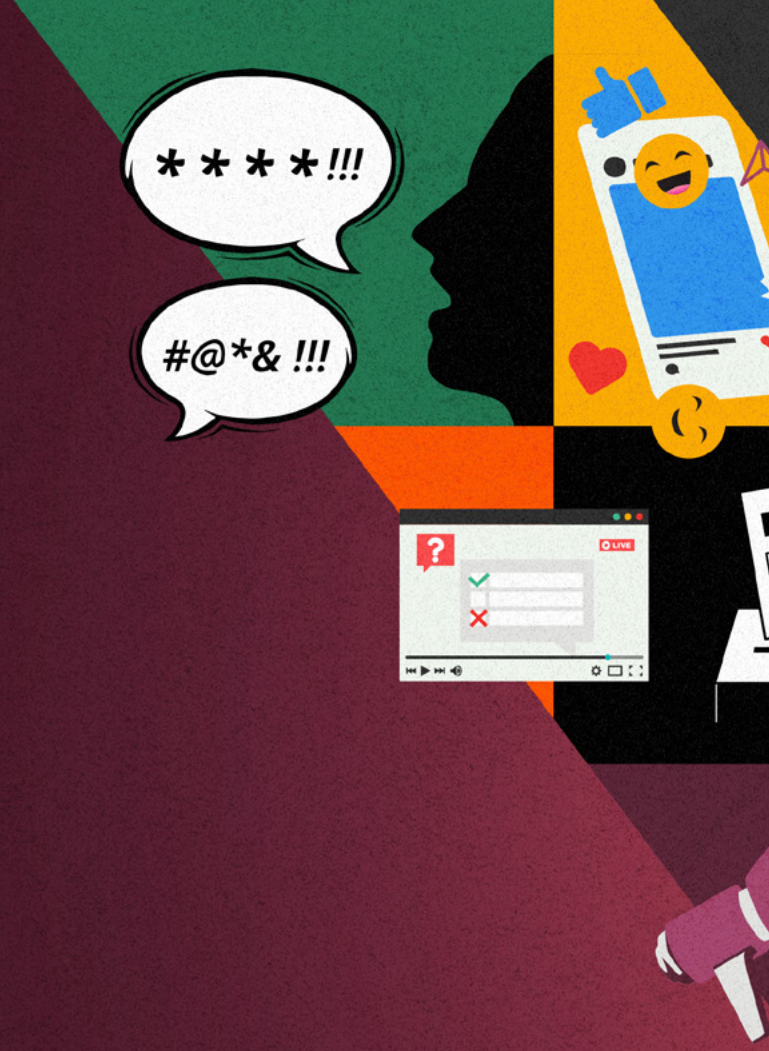
Annex I:

SMM glossary and hate speech classifications

TERMS	DEFINITIONS
Cheapfake	Content altered by technologically lowlevel manipulation of audio-visual material (created with easily accessible software).
Deep fake	Content manipulated or created by technologically highly sophisticated manipulation of audio-visual media, using AI-driven technology.
Disinformation	False information that is deliberately created or disseminated with the express purpose to cause harm.
Misinformation	Incorrect or misleading information that is unintentionally or unknowingly disseminated
Hate speech	Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor
Online Violence	Acts of online violence may involve different types of harassment, violation of privacy, sexual abuse and sexual exploitation and bias offences against social groups or communities.
Synthetic Media	Data and media (audio, text, image, or video) artificially produced, manipulated and modified by automated means, especially through the use of AI algorithms.
Political content	Content that refers to a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome. This also includes ads and posts about political issues and debates

TERMS	DEFINITIONS
Manipulated information	Authentic or non-authentic content that is manipulated to deceive public opinion or a specific audience.
Misogyny	A behavior, position, or speech that has a clear hostility or contempt towards women It is feelings of hating women, or the belief that men are much better than women
Political polarization	A process by which the public opinion tends to divide itself according to the relative closeness of each person to a particular political current. moving political attitudes to the ideological extremes.
Violence against women	Any act of gender-based violence that results in, or is likely to result in, physical, sexual, or mental harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life.
Discrimination	Is the act of making unjustified distinctions between human beings based on the groups, classes, or other categories to which they are perceived to belong. People may be discriminated against on the basis of race, gender, age, religion, or sexual orientation, as well as other categories.
Incitement	Is a very dangerous form of speech, because it explicitly and deliberately aims at triggering discrimination, hostility and violence, which may also lead to or include terrorism or atrocity crimes.
Swearing and insulting	Is the infringement of another human's honor by whatsoever means of expression
Defamation	Any statement made by a person, whether verbal or printed, that causes harm to another person's reputation or character.
Sexual Harassment	Any unwelcome sexual act such as physical contact, sexual hints or comments about color, pornography and sexual requests, whether in word or deed. This act can be humiliating and can lead to health and security problems.
Abasement	Any saying, writing, drawing, picture, or any reference or expression that humalitis and degrades the other

TERMS	DEFINITIONS
Fake news	Purposefully crafted, sensational, emotionally charged, misleading or totally fabricated information that mimics the form of mainstream news and are not supported by facts
Rumors	Circulated informationon social media, claims that not fact-checked, can be may be true, partially true, or may be false.
Manipulation of emotions	Speech based on attacking or showing the opponent negatively with the intention of winning the emotions of the public.
Self-promotion speech	A speech aimed at mobilizing supporters and direct endorsement through inducing the ballot.
Sexual harassment	Exposure to sexual or sexual content, phrases, materials, and images: (sexual images and videos, words and expressions that include sexual suggestions and images, threats of a sexual nature, such as threats of rape)
Sexism	Demonstrating bias, stereotyping, or discrimination against women and their gender roles. A masculine discourse based on stereotyping and discrimination against women.
Violence based on form/age, educational level, social level, religion, race.	For example, posting pictures of women how they were and how they became in reference to ugliness and beauty, and may be by avoiding positive criticism and generalizing mistakes to women in general.
Cyberbullying	Repetitive aggressive behavior aimed at ridiculing or offending: masculine speech directed at a specific woman, aimed at abusing or ridiculing her.
Threats and intimidation	Death threats, harm.



DEMOCRACY
REPORTING
INTERNATIONAL

Democracy Reporting International (DRI) was founded in 2006 by an international group of experts on democratic governance and elections.

DRI works on research and analysis to direct engagement with partners on the ground to improve democratic structures and safeguards across the countries where we work.

Elbestraße 28/29 12045 Berlin, Germany
info@democracy-reporting.org
wordsmatter@democracy-reporting.org
www.democracy-reporting.org/