DEMOS

ENGENDERING HATE

THE CONTOURS OF STATE-ALIGNED GENDERED DISINFORMATION ONLINE

APPENDICES

ELLEN JUDSON
ASLI ATAY
ALEX KRASODOMSKI-JONES
ROSE LASKO-SKINNER
JOSH SMITH

OCTOBER 2020

CONTENTS

APPENDIX 1
FULL NETWORK ANALYSIS

Below, we will take a deep dive into the universe of Twitter users who follow one or more of our key accounts in Poland and the Philippines. In each case study, we will look at these networks in three ways:

1. The **overall shape** of the network, in particular any evidence of any close knit 'echo-chamber' groups of users who all follow each other.
2. The **distribution of known gendered disinformation - and in the Philippines, counterspeech - across the network**, showing how prevalent gendered disinformation and messaging which speaks up against violence against women across different groups of followers.
3. The **character of various groups** within the network, discussing the types of key accounts present, words used by, and density of gendered disinformation-sharing users within groups of followers.

NETWORKS: POLAND

Figure I shows the network of users whose tweets were collected for our Polish investigation. These users consist of: the key users identified by the country team; a 10% random sample of those who *followed* at least one of our 30 key users, and all users *followed* by the key users ('friends'). The graph below shows how these key users, their friends and audience, are connected by followership.[114]

The overall shape of the network

Figure I below shows a complete network of all Twitter users following at least one of our key users. Because these busy, sneeze-like network maps can often be more visually impressive than they are useful, it is worth taking some time below to properly explain how this graph was produced.

Figure I was drawn using a program called 'Gephi', which was supplied with the key users in our collection, and a list of all the users who followed them.[115] The program is then instructed to:

- map these users out, connecting key accounts (shown in green below) to their followers and friends (shown in grey, as tiny points) by thin grey lines.
- These lines are drawn to be as short as possible, so that each user ends up close to those they follow.[116]

114. These graphs were built using Gephi, an excellent, free piece of open source software available from https://gephi.org/, and its foundational paper can be found at Bastian M., and others. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media, 2009.
115. Along with a few other stats, such as the number of Tweets each user had sent containing SAGD - we'll come to that later.
116. Note that we're not drawing lines for connections between two grey dots - we only know which key accounts a user follows, so non-key accounts following each other don't show up here.

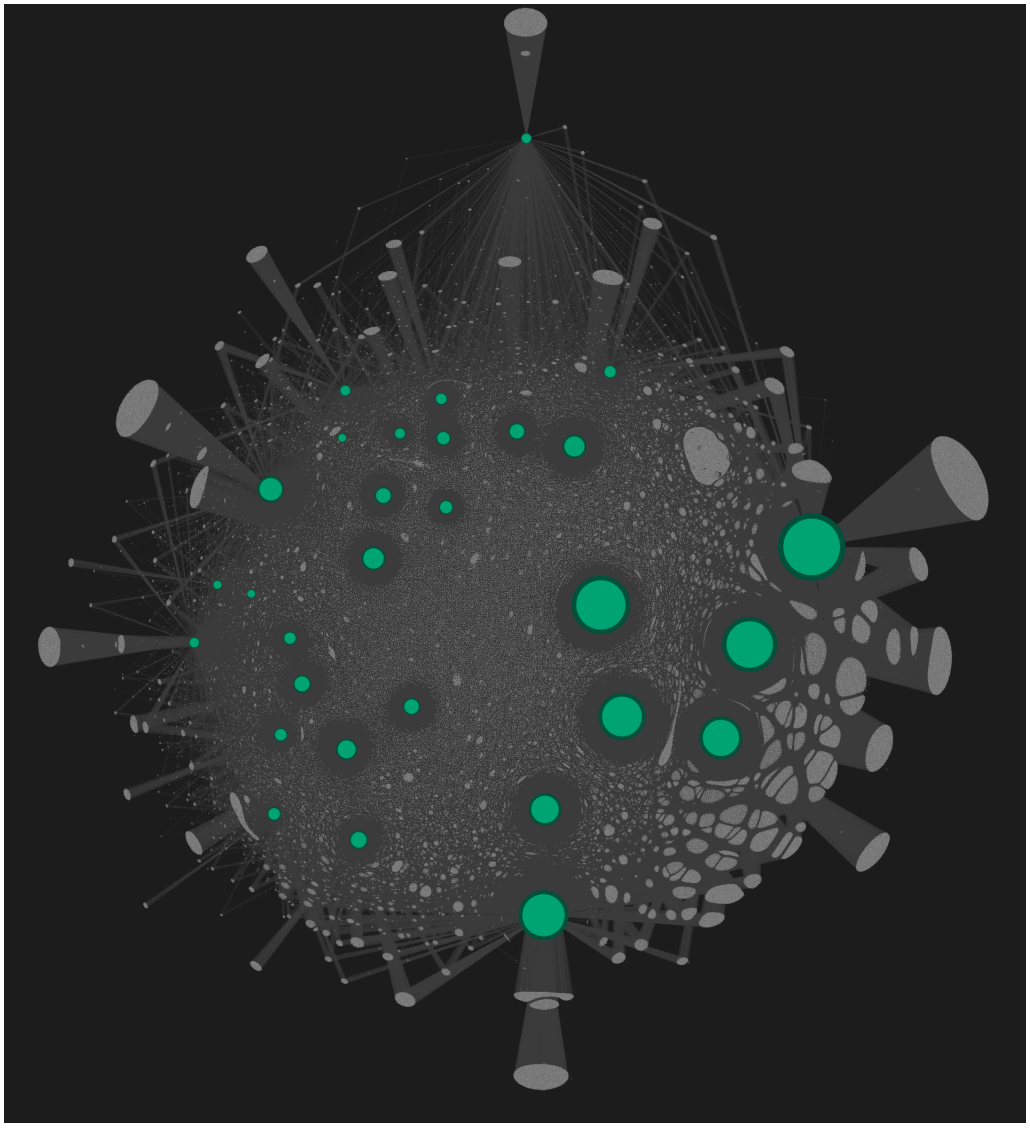


FIGURE I.
USERS FOLLOWING OR FOLLOWED BY POLISH SEED ACCOUNTS

The large green circles, (or 'nodes') above, showing key users, are sized by followership.

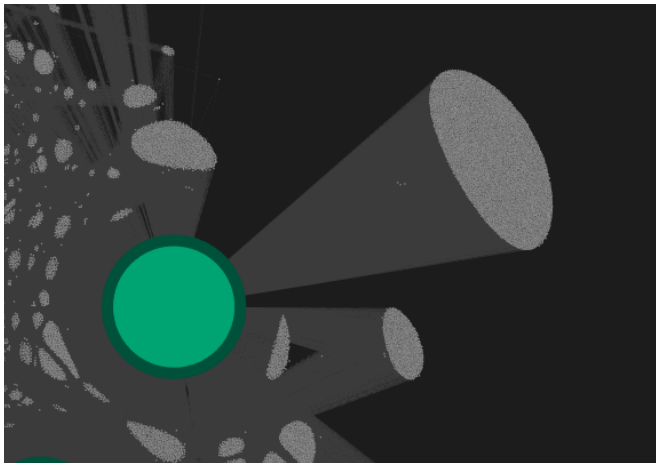
The more followers that user has, the larger the node.

Those ovoid grey clumps you can see at the bottom right of the graph are composed of hundreds of tiny dots, each representing a follower of a key user. Altogether, this graph represents our entire sample set - the users whose Tweets we collected and analysed for gendered disinformation.

In terms of overall shape, this graph is remarkably circular. With the exception of one unusual green node rocketing away from the top, belonging to a parody account of a broadcaster, each of the key users here are nestled into a single, coherent, circular group. This in contrast to the more spread out graph seen in the case of the Philippines below.

As we will see, different areas of this circle have different characters, and this affects their likelihood to share gendered disinformation. However, in the Polish case, these groups overlap, each with the other; they are well connected by users, (those in the centre of the graph,) who follow multiple key users. This has an effect on the social ecosystem - messages which are widely shared by one group of followers are more likely to be passed on amongst the circle, raising the possibility that gendered disinformation - as well as counterspeech - could more easily be spread to more users. We will see that this is in fact the case for this network below.

Another striking feature of Figure I, above, are the ovoid blobs which appear next to the large, popular accounts on the right - an example of which is shown here. These represent groups of users who all follow a single key user. For example, the large group to the left contains 9,244 users following a member of the European Parliament, but who don't follow any of our other key accounts. This pattern is repeated along the right side of the lower right half of the graph, which is thick with large, popular accounts.



These ovals indicate that, in addition to the well-connected users who have caused the graph's circularity, the ecosystem also contains users who typically follow only one or two accounts - in our case, primarily prominent politicians or academics - but are not connected to other key users with smaller followings. As we move to the left of the graph, this pattern vanishes and tends towards chaos - suggesting that the key accounts to the top left have both smaller followings, but are also more densely connected.

Known gendered disinformation in the Polish Ecosystem

131 of the users shown in Figure I were discovered, through sampling, to have sent tweets consisting of gendered disinformation.[117] These accounts are shown in Figure II. Users sending gendered disinformation are coloured and sized according to the number of relevant tweets in our sample; smaller yellow nodes sent a single tweet, larger red nodes sent more. Some of the largest, reddest nodes here are members of our key groups - the two accounts which sent the highest number of relevant tweets are both key accounts, with nine tweets each picked up in our sample. If that sounds like a small number, remember that this is based on only a small random sample of a few hundred tweets, containing specific keywords.

117. Tweets sent from the users in Figure I were filtered by keywords related to gendered disinformation, as above, and annotated for the presence of gendered disinformation.

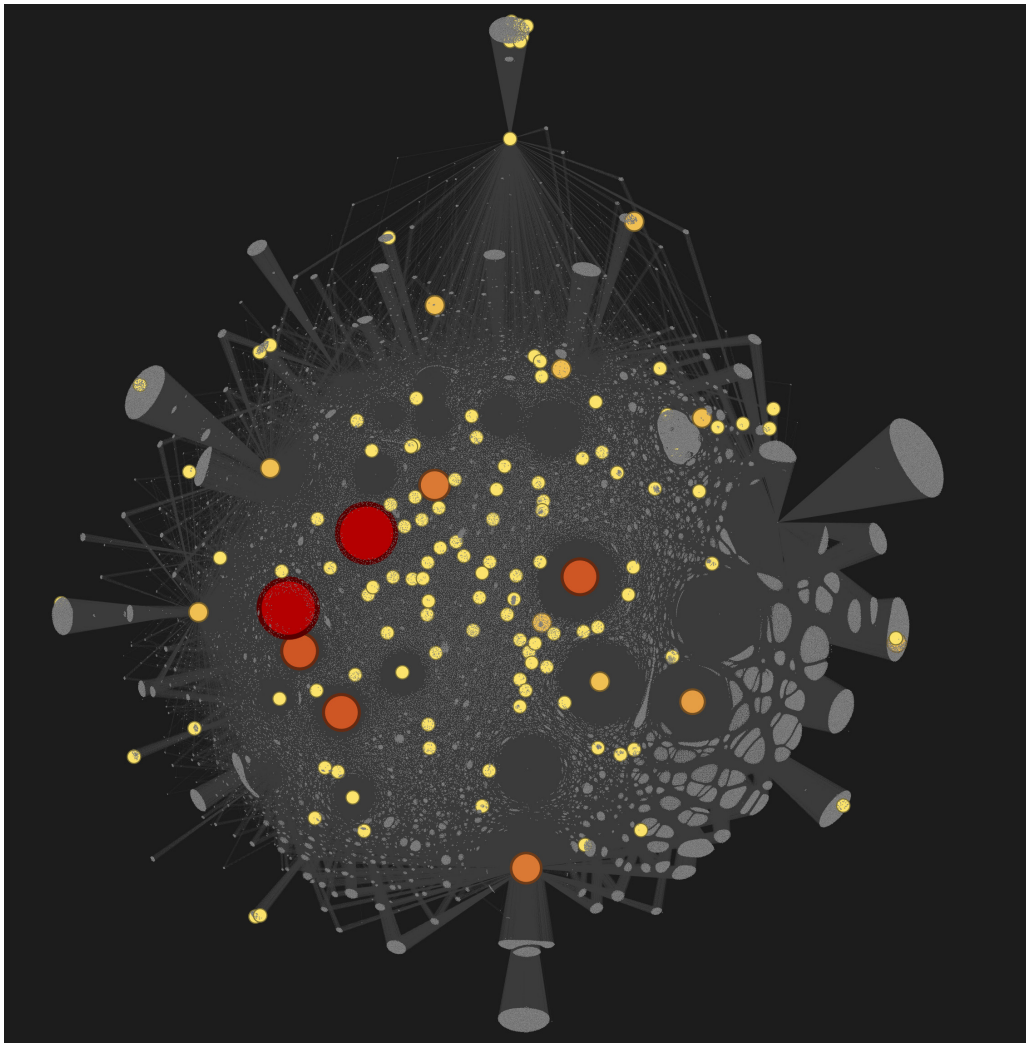


FIGURE II.
POLAND - USERS COLOURED AND SIZED BY DISINFORMATION

As a result of this sampling, the view above will greatly underrepresent the scale of disinformation being shared amongst our network. However, Figure II shows us how disinformation is likely to be distributed across the Polish ecosystem. Strikingly, the coloured points above seem to be scattered relatively evenly, suggesting that gendered disinformation is shared at a low level amongst most of the network. The exception to this are followers of the large political accounts to the right, whose sizable clumps of followers contain few tweets labelled as gendered disinformation even when the key user which they are following has sent relevant tweets. This could be an artefact of our sampling, but might also suggest that the majority of followers of prominent political accounts are less likely to author or reshare disinformation than those who follow multiple smaller accounts picked up in our sample, and are situated around the centre of the graph.

The character of groups within the network

Figure II suggests that, clustered around the two large red nodes on the left, is a group of users who are particularly likely to share or author gendered disinformation. To see whether this group could be identified and described, Gephi was asked to divide the graph into six groups, represented by different colours in Figure III below.[118] Put simply, accounts were grouped by how closely connected they are.

Each account is more likely to be connected to nodes of the same colour than it is to be connected to a node of another colour.[119]

118. This is known as 'modularity clustering', and it is one of the software's most interesting (and most abused) functions.
119. For a full description of this, see Blondel, VD. and others. Fast Unfolding of Communities in Large Networks. Journal of Statistical Mechanics: Theory and Experiment 2008.10, 2008: P10008. Crossref. Web.

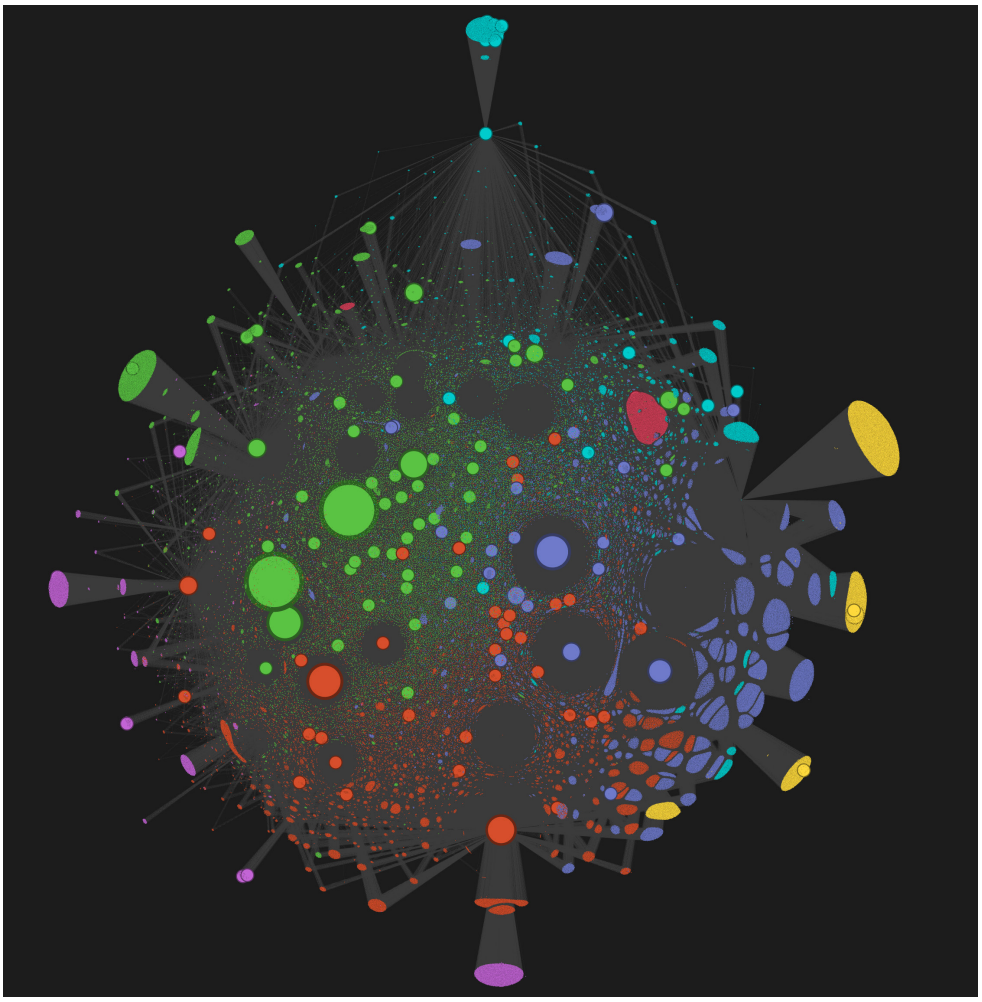


FIGURE III.
POLAND - TWITTER USERS COLOURED BY MODULARITY

In this graph, nodes are still sized by the amount of gendered disinformation shared - so the large accounts have sent at least one gendered disinformation tweet. This view of the network seems to show a higher number of relevant users in the green cluster, and fewer coloured blue. Table I shows this in more detail:

Cluster colour	No. of users	% of total users	Number of GD users	% of total GD users
TOTAL	**147,843**	**100%**	**131**	**100%**
Dark Blue	42,128	29%	22	17%
Green	32,645	22%	49	37%
Orange	27,666	19%	35	27%
Gold	16,711	11%	3	2%
Light Blue	14,896	10%	15	11%
Pink	7,878	5%	5	4%
Scarlet	5,919	4%	2	2%

TABLE I.
SIZE OF EACH COLOURED CLUSTER, WITH % OF 'GD' USERS SENDING AT LEAST ONE GENDERED DISINFORMATION TWEET

Table I suggests that there is a higher concentration of users sharing gendered disinformation in certain clusters. While the dense green group, identified as containing people relevant to political and academic life and the media, accounts for 22% of total users, but it contains 37% of 'relevant' users - those who shared content to the left of the graph accounts for 22% of total users, it contains 37% of 'relevant' users. In contrast, the large blue cluster, formed primarily around political accounts, contains 29% of the total userbase, but only 17% of relevant users - under half as many. These relevant blue users tend to be positioned towards the centre of the graph, and by implication connected to a higher number of key accounts. The orange cluster sits between the two both in terms of its position on the graph, and its saturation of relevant accounts.

The data presented above suggests a strategy for working out who is likely to spread gendered disinformation on Twitter, at least in the Polish online ecosystem.

Rather than focusing on high-profile, mainstream political accounts known to share disinformation, attention should be paid to smaller, denser networks of users. While each of the 'ringleader' accounts may have a relatively small followership, the above suggests that their audience is more likely to follow other relevant accounts, and more likely to have shared relevant material.

NETWORKS - THE PHILIPPINES

The overall shape of the network

Figure IV shows the network of users following any of our key users in the Philippines. As in Figure I, key accounts are shown as large green points, sized by the number of people who follow them, and their followers are shown as small grey dots. As in the Polish case, we can see that huge groups of users follow only one of our accounts, causing that distinctive ovoid 'hat' shape. Unlike the Polish example, however, this graph is much more spread out, and less clumped.[120] In particular, there are groups of key accounts which are spaced distantly from each other, with users in closely knit groups much less likely to follow relevant users from other groups.

This is important - as we will see below, it has an effect on the prevalence of gendered disinformation across the network.

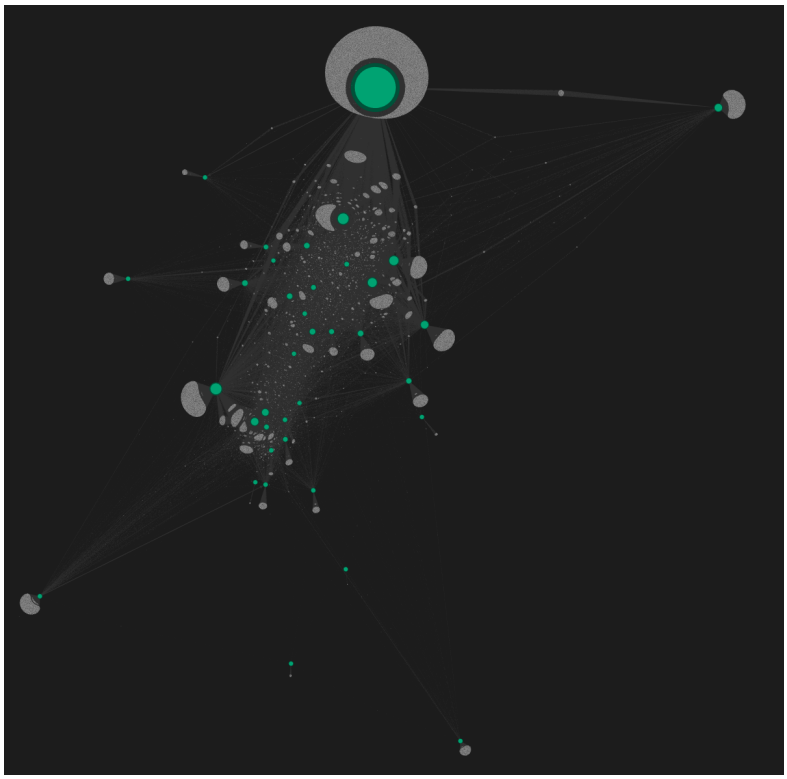


FIGURE IV.
PHILIPPINES - KEY USERS AND THEIR FOLLOWERS

120. This is despite using precisely the same rendering algorithm - the process within Gephi which determines where each node should be placed - for each graph.

Known gendered disinformation and counterspeech in the Philippines Ecosystem

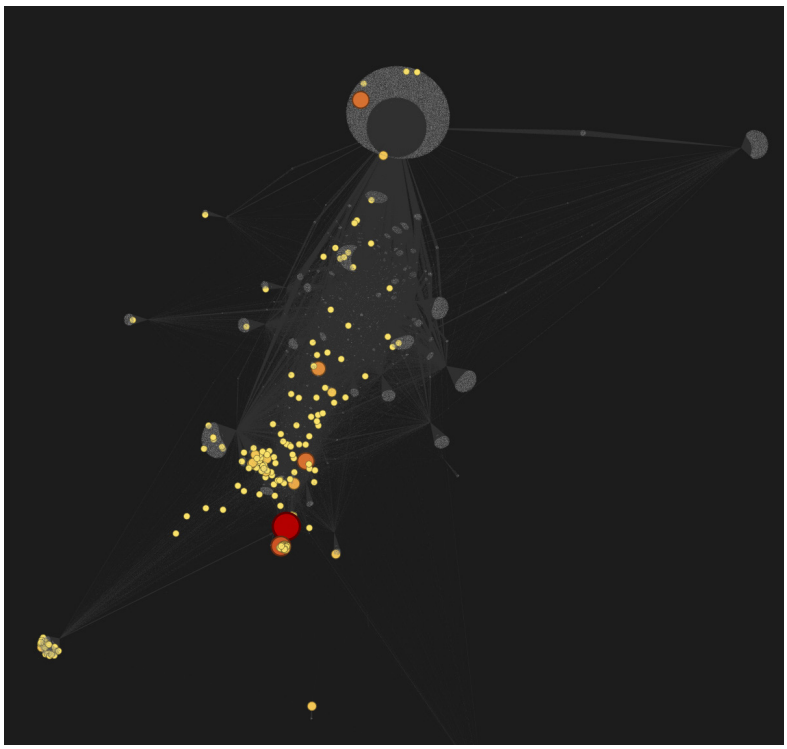


FIGURE V.i.
PHILIPPINES -
USERS SHARING GD

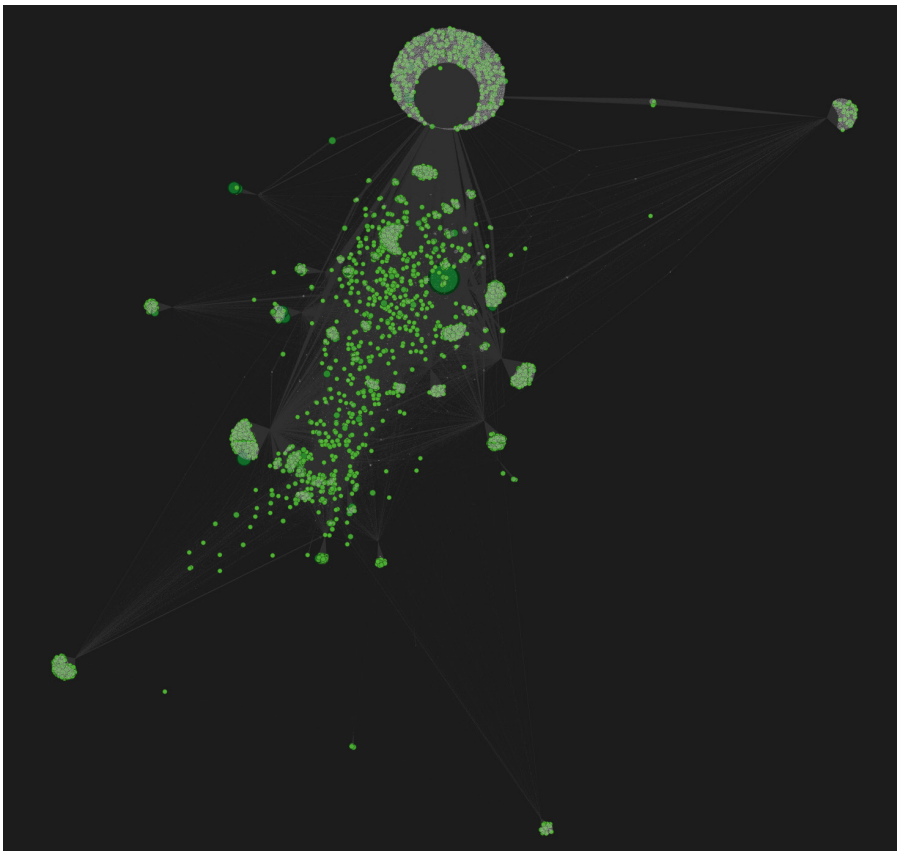


FIGURE V.ii.
PHILIPPINES -
USERS SHARING
COUNTERSPEECH

Figure V.i shows users coloured by yellow to red by the amount of known gendered disinformation they have sent - as in the Polish example, these gendered disinformation tweets were discovered through a manually coded sample identified through keyword filtering. Notably, and unlike in the Polish case, users sharing gendered disinformation tend to be concentrated in one area of the network - a group separated from other key users towards the lower right of the graph. This may represent an echo chamber - users who are more likely to follow others in the group than other key users elsewhere on the graph, and who, judging by this figure, are likely to see, send and share high quantities of gendered disinformation.

In the Philippines, unlike Poland, researchers were also able to train a 'counterspeech' classifier on tweets sent by those following or followed by key users' accounts, which used an NLP algorithm to determine whether or not a tweet was likely to consist of counterspeech. This process is described in detail below. Figure V.ii maps those counterspeech tweets on our network, with users who sent at least one tweet labelled as counterspeech shown in green. The difference to the gendered disinformation graph is striking, in that it is distributed across the entire network, including in areas which see high levels of gendered disinformation. This could be a heartening indication, suggesting that gendered disinformation on Twitter is being contested, to some extent, by the followers of those spreading it. It could also comprise a number of tweets misclassified as counterspeech (which we explore in more detail below.)

The character of groups within the network

As in the Polish example, Gephi was used to divide the networks into distinct groups. Each coloured user in Figure 6 is more likely to follow another account of the same colour than they are to follow an account of another colour.

Through inspecting the key accounts within each cluster, we can see in figure VI how different online political groups interact and overlap. The middle of the map shows official accounts of senators, which contain some groupings along party lines and show links to online media and influencers who support different political groups. Some of these accounts overlap with the most tightly clustered group - that of, mostly pro-Duterte cultural influencers some of whom have or have had links to state administration.

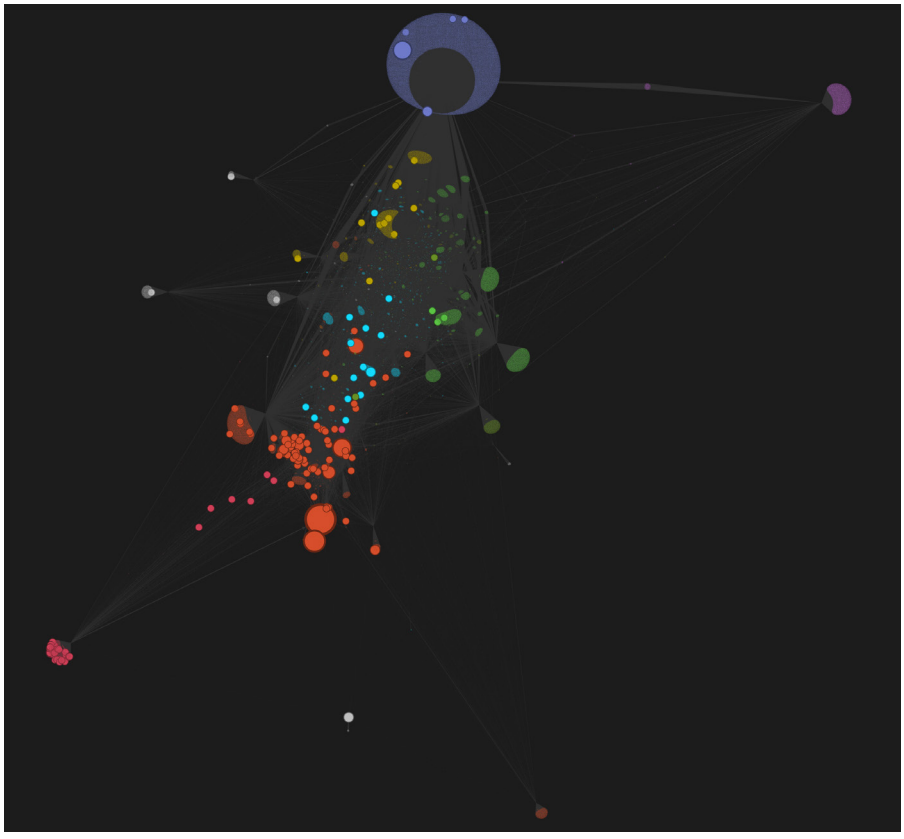


FIGURE VI.
PHILIPPINES - TWITTER USERS GROUPED BY MODULARITY

Most interesting here are three clusters in the centre of the graph. The ‘gold’ accounts are related to politicians who are known to speak out against gendered harassment, gendered disinformation and abuse, as well as supportive human rights activists. Lying right under this cluster (suggesting they share followers) is a ‘light blue’ group, comprised mostly of politicians, though also including media accounts and personalities. Some of these are linked with the government and some are critical of the government.

The 'orange' accounts are commonly associated with cultural influencers and bloggers, mostly pro-Duterte and/or critical of the opposition. These accounts show the 'revolving door' that exists between positions of political power and government authority, and those who are influential in online media and cultural discourse: some accounts are of those who have held both political and media roles at different times. Others are organisations, private individuals who comment on political affairs, and online political communities. Some are based overseas. This provides some useful context for the distribution of gendered disinformation, seen in V.i above, which seems to be concentrated in the orange, pro-Duterte cluster. This can be seen more clearly by looking at how many users across the various groups are responsible for sharing counterspeech and gendered disinformation.

Cluster colour	No. of users	% of total users	Number of GD users	% of total GD users	Number of counterspeech users	% of total counterspeech users
TOTAL	**166,947**	**100%**	**181**	**100%**	**3049**	
Dark Blue	79,286	47%	5	3%	319	10.46%
Green	24,125	14%	3	2%	481	15.78%
Orange	22,535	13%	110	61%	702	23.02%
Gold	11,899	7%	12	7%	631	20.70%
Light Blue	8,325	5%	14	8%	363	11.91%
Pink	7,841	5%	0	0%	30	0.98%
Scarlet	4,481	3%	31	17%	241	7.90%

TABLE II.
SIZE OF EACH COLOURED CLUSTER, WITH % OF ‘RELEVANT’ USERS SENDING AT LEAST ONE GENDERED DISINFORMATION OR COUNTERSPEECH TWEET

These numbers show a clear difference between groups. The orange, pro-Duterte group makes up a mere 13% of users on the graph, but account for 61% of users sharing gendered disinformation - though this cluster also accounts for 23% of users sharing counterspeech. In contrast, the gold cluster, key users following politicians speaking out against gendered harassment, accounts for 7% of the population and of gendered disinformation - but 21% of counterspeech. The dark blue group - the ‘hat’ at the top of the graph - makes up nearly half of followers but accounts for only 3% of gendered disinformation sharers and 11% of counterspeech sharers, strengthening the theory that these users are apolitical and unlikely to connected to the broader conversation underway in the rest of the network.

The ‘counterspeech’ seen in an otherwise gendered disinformation-heavy cluster may also be an artifact of misclassification. To investigate this, we examined language appearing in counterspeech, and tweets in general, shared by the orange and gold groups. In the ‘orange’ group, the group of

most interest in analysing gendered disinformation, hashtags shared often focused on undermining women politicians, and also on supporting the government. "#leniresign" was used 10 times, "#lenifakenews" six times, "#fakevprobredo" four times, and "#protectourpresident" (in conjunction with other hashtags) 10 times. The conjunction of "#lenipowergrabber" and "#protectprrd" five times in the dataset implies a connection being made between undermining women in politics in the ways set out above, with the 'protection' of the state. Interestingly, this group also used "#takebackthetech" six times - a hashtag more commonly associated with counterspeech.

In the 'gold' group of accounts, hashtags focused on speaking up against violence against women ("#metoo", six times; "#violenceagainstwomen", four times; and "#endrapeculture ph, 12 times), as well as critiquing the state.

The hashtag usage of these clusters alone support our identification of two different groups of particular interest: those who are broadly in support of or aligned with the state, who also join in with online conversations undermining prominent women in politics; and those who are further from the state and engage more clearly in counterspeech. The separation of these groups is by no means absolute, but does appear as a pattern also in the word frequency of tweets shared by users in these groups (below is a selection of the most relevant terms to gendered disinformation and counterspeech). We can see that similar terms appear across all three groups, but that the terms more associated with SAGD (boba, presstitutes, saba) appear more often in the 'orange' group, and the terms more associated with counterspeech (misogyny, sexual violence) appear more often in the 'gold' group, with the 'blue' group as a middle ground.[121]

	Orange	Blue	Gold
boba	111	20	27
duterte	67	40	98
presstitutes	15	0	0
rappler	10	0	0
slut shaming	6	5	14
bitch	73	0	0
saba	34	13	12
rape	405	305	527
slut	12	6	27
misogyny	0	0	7
rape jokes	20	16	49
sexual violence	0	0	10
sexual harassment	0	0	13
duterte supporters	0	10	27

Some terms appear across all three at similar frequencies - sometimes this is likely to be because something was the topic of broad discussion (e.g. "rape jokes" was widely discussed on all sides after Duterte made such a comment). "Rape" was used extremely often. Given the low incidence of physical threats we observed, most of these are likely discussions of news stories of rape. Some of these will have been used in contexts of condemning rape explicitly, which, as discussed, happens across the political spectrum: from some sides, to demand women's rights, from other sides, as described in Rule 5 to justify hardline policies and a 'strongman' approach. These tweets would likely have been classified as 'counterspeech' on the basis that they condemned violence against women, whichever side they were on.

121. "Bitch" appearing on its own was found only in the orange group, but phrases which included the term e.g. "yes bitch" were found in other groups.

APPENDIX 2
FULL METHODOLOGY

Working with the National Democratic Institute in the USA, we identified two 'trial' countries to investigate: Poland and the Philippines. These countries were selected based on: the prevalence of disinformation and/or gendered abuse; political attitudes towards gendered issues; levels of online political discourse across different platforms; where there were strong connections with in-country partners to guide the analysis; and language capabilities of the research team and how they mapped onto languages used online in each country.

In-country partners in Poland and the Philippines completed a) a political network mapping and b) a gendered lexicon. The political network mapping identified individuals who "can and do influence the terms and tenor of political information online", who operate within or are linked to political networks. The categories of relevant individuals included elected officials, their staff and operatives, state-affiliated actors and bureaucrats, families of state-affiliated officials, and religious, business and cultural leaders. As state-aligned disinformation was of primary interest, however, the focus was on state-aligned individuals - though not exclusively. This mapping gave us set of core individuals whose online activity we could investigate to see if gendered disinformation was pervasive within it, and if so, examine which kind of influential figures (from directly state-employed to more widely state-aligned, or even state-opposed) were most commonly engaging in creating or sharing it.

From the Poland in-country team, 30 relevant individuals were identified, mostly people involved in political life, but who fell into a variety of political or cultural categories, some of which overlapped with each other. These formed the initial usernames we examined tweets from.

From the Philippines in-country partners, 69 relevant individuals were identified. However, many of these primarily used Facebook, and Twitter accounts could not be located for all of them. The initial composition of usernames from the Philippines was primarily cultural leaders or relevant influencers, rather than politicians, as "very few ranking officials in the Philippines manage their own accounts", and "for many of these personalities, we do see some support on the part of government. For instance, they are in official government functions, they are featured in government channels, etc." Of these, we identified 23 related Twitter accounts. As this was a small sample, we supplemented the list with 19 Twitter accounts we identified belonging to politicians in the Philippines, giving 42 total usernames.

In-country partners also compiled lexicons of terms: "phrases will either target individual women as political leaders in an effort to drive them out of politics, or they will use gender norms to: manipulate the views of men and women on women's leadership in general and/or of gender-related issues; influence specific political outcomes." These terms were either commonly used in online gendered violence to harass and abuse people and/or in political discourse (slogans, hashtags etc.). These were categorised according to the NDI's typology of online violence against women in politics (VAW-P) into terms commonly used for the purposes of insults and hate speech, embarassment and reputational risk, physical threats, and sexualised distortion (or purely political if not related to VAW-P).

The Polish lexicon included 66 words or phrases, including: 15 used for the purposes of embarrassment and reputational risk; 25 used as insults or hate speech, two in making physical

threats and 24 used in sexualised distortion. 43 words/phrases were also provided to form a Political lexicon, which included words and phrases relevant to politics and gendered political issues.

Contextual information was included about features of the use of the term, including: whether the term was always pejorative; whether the term was only relevant in the region; whether the term was in constant use; and whether the term had been reclaimed or not. These were used by the research team to better understand the context and meaning of the terms and highlighting which words might be more useful than others.

These features were not, however, used in filtering the data as they were not specific to SAGD (as opposed to gendered language in general): this was done based on analysts' observation of which terms occurred within the dataset in which contexts. Generally, the terms which were identified as both specifically political and gendered were the most useful in identifying SAGD.

The Philippines lexicon included 34 words or phrases that were related to gender-adjacent harassment, and five related to gendered political language, as well as two specific instances of gendered political language (not used for searches due to their specificity).

Data Collection

Tweets were collected using Twitter's streaming API. This collection and analysis was done using Method52, a suite of tools for collecting and analysing large free-text datasets developed by Demos in partnership with the University of Sussex.

We first collected tweets from the seed users identified through the political mapping phase, collecting tweets since January 1 2019. As the datasets were relatively small, we also collected the first page of tweets sent from accounts followed by the seed users ('friends') and from a random 10% sample of accounts which followed the seed users ('followers'), due to the high volume of followers.

From 42 initial usernames from the Philippines, we collected 1,523,301 followers of our initial set and 16,057 'friends' (accounts followed by the initial sets). We collected 52,200 tweets from our initial username set, 2,598,275 tweets from their friends, and 5,521,987 tweets from a random 10% (152,903) sample of their followers.

From 30 initial usernames from Poland, we collected 1,626,140 followers of our initial set and 37,373 'friends' (accounts followed by the initial sets). We collected 46,525 tweets from our initial username set, 6,223,963 tweets from their friends, and 7,157,983 tweets from a random 10% sample (134,534) of their followers.

Data Filtering

We then filtered the tweets so that those remaining were those which contained a term or phrase from the gendered lexicon.

The seed user tweets from the Philippines were filtered for English language tweets. The remaining dataset contained 148 tweets.

The seed user tweets from Poland were filtered to remove only matches with the Political Lexicon to ensure that remaining results had a gendered element. The remaining dataset contained 149 tweets.

On examination of the data from the followers and friends, it was apparent that there were a lot of irrelevant matches from where a gendered slur had been used in a way that was not relevant to gendered disinformation - terms such as "kurwa", "dupa" and "cholera" in Polish, and "puta", "sexy" "bitch" and "slut" in the Philippines data which are used in a wide variety of contexts.

In order to filter out these terms which were in common usage but were not relevant, we used a conditional filter such that tweets containing those terms which were likely (judged by analysts on the basis of the lexicons and the data examined so far) to be 'noisy' result in irrelevant results would only be included if a 'significant' term was also present in the tweet. A list of 'significant' terms was compiled by analysts on the basis of the data examined so far, the lexicons, and the background literature review, and included terms associated with likely counterspeech, names of likely targets of gendered disinformation, and other terms from the initial lexicon. All tweets containing gendered terms other than the 'noisy' ones were included. The Philippines dataset was then filtered for English language.

This resulted in a set of 4,316 tweets from followers and friends of the Philippines usernames and 17,640 tweets from followers and friends of the Polish usernames.

Automated Data Analysis

Classifiers were built on the Philippines dataset initially.

Firstly, the seed tweets from the Philippines were coded manually by analysts, using the NDI typology as a starting point and seeing what other trends emerged from the data, with input from the in-country team where clarity on context or meaning was needed. This stage highlighted counterspeech in particular as a significant category worthy of further exploration. The categories identified were: gendered threats, gendered embarrassment, sexualised distortion, gendered insults and hate speech, victim blaming, general criticism of prominent women, reporting personal gendered abuse, reporting gendered abuse/violence against another in a critical way, religion and abuse, criticising rape jokes, defending rape jokes, pro-state leader speech, counterspeech, irrelevant use of gendered slurs, other, and a mismatch of a term (for instance where "Lady Gaga" was picked up because "gaga" was in the lexicon). This was then used to guide the next stages of the analysis.

A classifier was built and trained to identify counterspeech, a significant portion of the dataset, where counterspeech was understood as speech which was critical of or called out gendered disinformation, abuse, harassment or violence against women, taken broadly (e.g. condeming rape as well as specifically calling out online GBV or VAW-P).

Label	Precision	Recall	FB1	Labelled
relevant	0.736	0.750	0.743	114
irrelevant	0.934	0.929	0.932	145
Acc.			0.892	

This classifier was then used on the relevant 'friends and followers' tweets to filter out counterspeech. The remaining dataset (which was the data most likely to be relevant to gendered disinformation) contained 3,186 tweets.

To refine this data further, classifiers were built and trained to identify 1) gendered insults targeted at individuals specifically and 2) gendered disinformation in its broadest sense, but did not reach the minimum levels of precision and recall to be usable to classify within the dataset more widely, necessitating further refinement of the dataset using filters to allow for manual classification.

This was due primarily to two reasons: firstly, the low proportion of tweets within those collected which were relevant to gendered disinformation, meaning that in a sample the number of relevant tweets was too small for a classifier to adequately learn from. Secondly, identifying whether a tweet was relevant or not, for a human analyst relied heavily on background knowledge of the context in which the tweet was written, either from background evidence reviews, input from the country teams, or further exploration around a specific tweet (for instance what picture it was shared alongside). Automated software using natural language processing, therefore, which had access only to the words and syntax being used in the tweet, struggled to make these identifications at an acceptable level of precision and recall.

1)

Label	Precision	Recall	FB1	Labelled
relevant	0.400	0.571	0.471	16
irrelevant	0.967	0.935	0.951	48
Acc.			0.910	

2)

Label	Precision	Recall	FB1	Labelled
relevant	0.350	0.824	0.491	101
irrelevant	0.950	0.687	0.797	112
Acc.			0.710	

Since these classifiers were unsuccessful, and as much of the relevant tweets being reviewed by analysts in the course of building and training classifiers was targeted at particular individuals or groups, it was decided to further refine the dataset into 'generic' and 'targeted' tweets, by filtering for names or phrases (identified through data examination, the lexicon and background literature review) which were associated with a specific target of gendered violence or disinformation.

Attempts to classify on the 'generic' tweets met with the same difficulties as formerly - namely, that the proportion of relevant to irrelevant tweets within a sample (13 to 87) was too low to successfully train a classifier. This confirmed that the amount of

irrelevant tweets in the dataset was significant, and so it was decided to make the focus the targeted tweets which were much more likely to be relevant. There were 590 of these, and so they were coded manually by analysts, using the NDI categories as a basis and expanding the categories where necessary.

Given the difficulties building classifiers on the Philippines dataset, and the fact that similar phenomena were present in the Polish dataset (for instance, tweets which were difficult for a human to decidedly categorise into a particular classification, and so would have been significantly more challenging for automated software), it was decided to approach the Polish dataset in a similar way. A list of likely targets of gendered disinformation was supplied by the in-country team, and this used to filter the data, which resulted in 269 tweets. This was then manually coded by a Polish-speaking analyst.

Manual Data Analysis

Tweets from the seed users and the followers and friends were analysed qualitatively to identify a) common stories which were being told about the targets of the gendered disinformation b) strategies, tactics or patterns on display in the dataset. Researchers coded the tweets in the Philippines dataset manually into the following categories: gendered insults and hate speech, embarrassment and reputational risk, physical threats, sexualised distortion, counterspeech, irrelevant, and relevant news (this was news shared which was relevant to the phenomenon of gendered disinformation without necessarily being an instance of the phenomenon itself, but treated as similar since discussion of these campaigns can also contribute to the phenomenon spreading). The data was also examined to identify how many unique users were present in the relevant data, their unique locations, the range of their friends and followers, and their status as verified Twitter users. The initial political mapping was used again to identify which kinds of users were present in the final dataset. The Poland data was also coded into relevant/irrelevant tweets.

In total, 176 Polish tweets from seed users, their followers and friends were identified as relevant to gendered disinformation. 290 tweets from the Philippines were identified as relevant to gendered disinformation.

Network Maps

A simple Python script was written to annotate a list of follower relationships between users with labels showing which of those users had been found through the above process of analysis to have shared a Tweet relevant to misinformation. This data was then processed and displayed using Gephi.

Method52's inbuilt URL expander was used to expand the URLs which had been shared in tweets identified as relevant to gendered disinformation or counterspeech. Using the URLs from the Tweets classified as relevant, we reviewed the types of media shared by original users and the friends and followers.

LIMITATIONS, CHALLENGES AND METHODOLOGY RECOMMENDATIONS

Our recommendations for refining the methodology are as follows:

1. A larger or more diverse pool of initial users and more extensive lexicon would have enabled more data to be gathered. This may have meant higher proportions of gendered disinformation in the dataset which would have allowed a classifier to learn better. However, given that across the two countries the proportion of gendered disinformation content to irrelevant content was small, it is possible that this would have been the same in a larger set. It should be noted that the user and lexicon are determinative of the kind of gendered disinformation that is uncovered, and hence a broader list would also enable more specific kinds and themes to be detected, as the final dataset of relevant tweets used in this research was relatively small.

2. Human analysts are superior to automated software in detecting gendered disinformation, and require contextual and language capabilities in order to identify it most accurately. The successfully built classifier still showed errors in both precision and recall. Although no classifier will be perfect, the difficulties in building other classifiers to process the data, given the nuanced contours of these categories, suggest that attempts to counter gendered disinformation on online platforms which rely on the use of automated software or human content moderators who are not familiar with the local context are unlikely to succeed. They may over-moderate (for instance, by removing any tweet with a gendered slur in it even though many of those uses are irrelevant to gendered abuse or disinformation) or under-moderate (for instance, by relying on an automated system which cannot precisely distinguish between harassing speech and counterspeech given the similarity of the topics discussed and so errs on the permissive side). The need for contextual information and understanding is often cited as a drawback of relying on automation for content moderation of other kinds of harmful speech.[122,123] In a similar vein, a limitation of this study was that due to language capabilities of the research team, of the tweets in the Philippines dataset, only tweets in English could be examined.

3. Gendered slurs have become commonplace in discussions of all manner of topics, both positively and negatively, in casual conversation or about irrelevant topics. This is itself a reflection of the gendered nature of current discourse - where terms which have their origin in the abuse of women become common vernacular (in a non-reclaimed way), this itself can change the nature of online discourse. However, for the purposes of this

122. ee e.g. Gomes, A. and others. Drag queens and Artificial Intelligence: should computers decide what is 'toxic' on the internet? Internet Lab, 2019. Available at https://www.internetlab.org.br/en/freedom-of-expression/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/ [accessed 15 May 20]
123. Stecklow, S. Why Facebook is losing the war on hate speech in Myanmar. Reuters, 2018. Available at https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/ [accessed 15 May 2020]

project the focus was specifically on gendered disinformation and so these results had to be excluded. To counter this, and remove this 'noise' from the dataset, we took two approaches, which in combination succeeded in producing a highly relevant dataset. We first removed tweets, where the only word from the lexicon which it contained was one we had seen very often appear in irrelevant tweets, such as "sexy". Tweets where this word appeared alongside another word likely to be relevant to gendered disinformation, such as "presstitute", were included. We then filtered the data so only tweets mentioning names and variants of names of known SAGD targets were included to produce a more highly relevant dataset. These were broadly successful methods. However, even after the keyword filtering was complete, human analysts still detected irrelevant results in the dataset, highlighting the importance of contextual analysis of data gathered.

4. A limitation, however, of focusing on targets was that it limited the opportunity to discover new targets of gendered disinformation, others than those which were already known. However, we were able to expand our target list through initial scoping of the data which revealed prominent women who were frequently mentioned or targeted, and we recommend this step as a means to expanding the final dataset to be examined.

5. We were able to examine other platforms using Twitter as a stepping-off point to identify where other discussion was taking place. However, due to the lack of open APIs from many other platforms, we were not able to collect data programmatically from these spaces and relied instead on manual reviews and background context provided by the country teams. We would recommend that platforms make more data generally available to researchers to assist in this form of research, which otherwise can be skewed towards Twitter for technical rather than social reasons.

6. A limitation of the method was also that it cannot be seen to be representative of gendered disinformation - we identified a snapshot of gendered disinformation and relevant material, and not by any means the full spectrum or ecosystem of it.

7. Identifying the tweets relevant to gendered disinformation can also not be taken as identifying precisely the originators and actors of gendered disinformation. This is always a challenge - particularly in contexts such as the Philippines where networks of gendered disinformation actors can be very informal - rather than tight central control of campaigns, actors may be "loosely" connected, control decentralised and the separation between paid actors and the general public bridged by unpaid "grassroots intermediaries" such as those who run political fan pages.[124] Inferring intention and coordination from a dataset of tweets is very challenging, and as such here we have aimed to primarily identify content which aligned with the aims of state disinformation campaigns.

8. Identifying state-originating gendered disinformation was contingent upon having sufficient state actors in the original political mapping. In countries such as the Philippines where state actors have less of an online presence, supplementing this investigation with other methods may be productive in tying disinformation campaigns to a state.

COUNTRY-SPECIFIC METHODOLOGICAL CHALLENGES

Our partners from Poland and the Philippines highlighted specific challenges with creating a country-specific gendered lexicon and identifying relevant users for analysis. These are outlined below.

1. Slang and language are often gendered and nuanced. In Poland, "most swear words center on genitalia and sexual violence" and can be directed at both men and women. It is difficult to determine whether it is specifically gendered abuse or more conversational. The natural language processing tools similarly struggled to pick up the nuance of gendered misogyny online.

2. The regularity of synonyms in the Polish language in particular made it easy to miscategorise gendered slurs.

3. The nature and number of accounts and users on Twitter meant that the sheer number of

124. Ong, J.C. and Cabanes, J.V. Architects of Networked Disinformation. University of Leeds, 2018. Available at http://newtontechfordev.com/wp-content/uploads/2018/02/ARCHITECTS-OF-NETWORKED-DISINFORMATION-FULL-REPORT.pdf [accessed 12 March 2020]

accounts made mapping a network of relevant users challenging. The fact that a lot of gendered slurs come from anonymous accounts or accounts that had been closed made it challenging to pick them up. In the Philippines in particular "official accounts, politicians are generally more controlled" so the content from their accounts were less likely to include misogyny, instead it is mostly perpetuated by the president's supporters who echo their politicians' sentiments.

4. Misogyny shared in visual mediums such as memes, and which is embedded in the subtext of conversations rather than overt, is very common, but difficult to pick up with natural language processing tools.

Background: on Natural Language Processing and the NLP Classifier

Building algorithms to categorise and separate tweets responds to a general challenge of social media research: the data that is routinely produced and collected is too large to be manually read.

Natural language processing classifiers provide an analytical window into these kinds of datasets. They are trained by analysts on a given dataset to recognise the linguistic difference between different kinds of data, in this case between tweets. This training is conducted using a technology called 'Method 52', developed by the project team to allow non-technical analysts to train and use classifiers. Each classifier was built by using Method 52's web-based user interface to proceed through the following phases:

Phase 1: Definition of Categories The formal criteria explaining how tweets should be annotated is developed. Practically, this means that a small number of categories – between two and five – are defined. These will be the categories that the classifier will try to place each (and every) tweet within. The exact definition of the categories develops throughout the early interaction of the data. These categories are not arrived at a priori, but rather iteratively, informed by the researcher's interaction with the data – the researcher's idea of what comprises a category will often be challenged by the actual data itself, causing a redefinition of that category. This process ensures that the categories reflect the evidence, rather than the preconceptions or expectations of the analyst. This is consistent with a well-known sociological method called 'grounded theory'.

Phase 2: Creation of a Gold Standard Test Dataset This phase provides a source of truth against which the classifier performance is tested. A number of tweets (usually 100, but more are selected if the dataset is very large) are randomly selected to form a gold standard test set. These are manually coded into the categories defined during Phase 1. The tweets comprising this gold standard are then removed from the main dataset, and are not used to train the classifier.

Phase 3: Training This phase describes the process wherein training data is introduced into the statistical model, called 'mark up'. Through a process called 'active learning', each unlabelled tweet in the dataset is assessed by the classifier for the level of confidence it has that the tweet is in the correct category. The classifier selects the tweets with the lowest confidence score, and these are presented to the human analyst via a user interface of Method52. The analyst reads each tweet, and decides which of the pre-assigned categories (see Phase 1) that it should belong to. A small group of these (usually around 10) are submitted as training data, and the NLP model is recalculated. The NLP algorithm then looks for statistical correlations between the language used and the meaning expressed to arrive at a series of rules-based criteria, and presents the researcher with a new set of tweets which, under the recalculated model, it has low levels of confidence for.

Phase 4: Performance Review and Motivation The updated classifier is then used to classify each tweet within the gold standard test set. The decisions made by the classifier are compared with the decisions made (in Phase 2) by the human analyst. On the basis of this comparison, classifier performance statistics – 'recall', 'precision', and 'overall' (see 'assessment of classifiers', above) - are created and appraised by a human analyst.

Phase 5: Retraining Phase 3 and 4 are iterated until classifier performance ceases to increase. This state is called 'plateau', and, when reached, is considered the practical optimum performance that a classifier can reasonably reach. Plateau typically occurs within 200-300 annotated tweets, although it depends on the scenario: the more complex the task, the more training data that is required.

Phase 6: Processing When the classifier performance has plateaued, the NLP model is used to process all the remaining tweets in the dataset into the categories defined during Phase 1, using rules inferred from data the algorithm has been trained on. Processing creates a series of new data sets – one for each category of meaning – each containing the tweets considered by the model to most likely fall within that category.

Phase 7: Post Processing Analysis After tweets have been processed, the new datasets are often analysed and assessed using a variety of other techniques.

CLASSIFIER PERFORMANCE

No NLP classifier used on this scale will work perfectly, and a vital new coalface in this kind of research is to understand how well any given algorithm performs on various measures, and the implications of this performance for the research results. Each classifier trained and used for this paper was measured for accuracy. In each case, this was done by: 1. Randomly selecting 100 tweets to comprise a 'gold standard'. 2. Coding each of these tweets by hand, conducted by an analyst. 3. Coding each of these tweets using the classifier. 4. Comparing the results and recording whether the classifier got the same result as the analyst. There are three outcomes of this test. Each measures the ability of the classifier to make the same decisions as a human in a different way:

Recall

Recall is a measure of the correct selections that the classifier makes as a proportion of the total correct selections it could have made. If there were 10 relevant tweets in a dataset, and a relevancy classifier successfully picks eight of them, it has a recall score of 80%.

Precision

Precision is a measure of the correct selections the classifier makes as a proportion of all the selections it has made. If a relevancy classifier selects 10 tweets as relevant, and eight of them actually are indeed relevant, it has a precision score of 80%.

Overall – F SCORE

The 'overall' score combines measures of precision and recall to create one, overall measurement of performance for the classifier. All classifiers are a trade-off between recall and precision. Classifiers with a high recall score tend to be less precise, and vice versa.

DEMOS

PUBLISHED BY DEMOS OCTOBER 2020

15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK