

Module 10: Disinformation

Analysis

Kalina Bontcheva (k.bontcheva@sheffield.ac.uk)



This session will be recorded

Recorded video will be available after this session



TYPES OF INFORMATION DISORDER



Very difficult to distinguish □ mainly between MIS- and DISinformation

Information disorder theoretical framework (Wardle, 2017; Wardle & Derakshan, 2017)



4

Misinformation lifecycle?





The 6 Questions of Misinformation Analysis

- What is being spread?
- Where it spreads?
- How it spreads?
- Who is spreading it?
- Why it spreads?
- When it spreads?

5







Follow

Last photo taken of the shooter that is shooting YouTube HQ. Goes under the name "Samuel Rodriguez Hyde". Said to be one of DailyMotion's biggest personalities. Motivations unclear.



3:15 PM - 3 Apr 2018

12 Retweets 24 Likes



NOTORIOUS MASS MURDERER AND WHITE SUPREMACIST, SAM HYDE CONFIRMED AS SHOOTER AT YOUTUBE HQ #YoutubeHQShooting #YouTube #Shooting #MarchForOurLives ### #NeverAgain



https://www.buzzfeednews.com/article/ishmaeldaro/roundup-of-misinformation-on-youtube-shooting







WeVerify project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825297

Leave the E

The Trumpet of Amplification



General Architecture for Text Engineering

G

Source: <u>https://medium.com/1st-draft/5-lessons-for-reporting-in-an-age-of-disinformation-9d98f0441722</u>



Misinformation in search results



ferguson was lie ferguson was staged ferguson was staged ferguson was not about race ferguson was a thug ferguson was planned ferguson was he armed ferguson was a hoax ferguson was fake ferguson was not racist I'm Feeling Lucky

9

Report inappropriate predictions

Source: (Albright, 2018)



Same image, different disinformation narratives and languages

inews 4 24 marzo alle ore 10:12 · ©

ULTIM'ORA - Dottoressa e OSS aggredite in Pronto Soccorso -GUARDATE COSA FA QUESTO UOMO IIII

Condividi per Denunciare



1,4 mln visualizzazioni

Mi piace	Commenta	A Condividi	S -
🖞 🥪 😵 5,6 mila			Commenti più in vista
26.440 condivi	sioni		3,1 mila comment



La verdad detrás del bulo sobre la agresión de un musulmán en un centro médico

La historia, difundida en Facebook, fue protagonizada por un hombre ebrio y ocurrió en R



Imagen del vídeo difundido por Facebook Rusia quiere regular el uso drones tras varios ataque:

Le supuesta historia de un supuesto musulmán que agredió a un médico en un centro de salude español ha sobregasado los tres miliones de visitas en Faceboco. Pero solo la mitad del relato es verdadero: Es cierto que hubo una agresión en un centro de salud pero el protagonista era un ruso en estado de embriaguez y ocurrió en la localidad rusa de Nóvgorod.

En las imágenes, que Juan Manuel González Sosa publicó en su cuenta de Facebook el pasado 25 de marzo, se aprecia cómo un hombre zarandea y golpea a una persona vestida con una bata blanca. Segundos después aparece una

Hame page Politika History Education Science Health Urban Legends

APUNTATE A NUES

TE PUEDE INTERESAR

videojuego como prueba "irrefutable" de que EE UU

islamófobo que retuiteó Ti Turquíe, Siria y Rusia carg los planes de EE UU de cre foerza fronteriza kurdo-sir

con el ISIS

bases en Siria

#STARTMEAPP

WHAT? METHODOLOGY AND PRINCIPLES TEAM NOTICE CONTAC

Syrians refugee allegedly beat a doctor from Russia shared video



Credit: Denis Teyssou (AFP)



	Images	Maps	Shopping	More		Settings	Tools
oout	92,300 resu	lts (1.11 se	conds)				
	17	Imag 346	e size: 259				
2			200				
1	200	Find	other sizes of t	his image:			
	22	Find	other sizes of t zes - Small - La	his image: arge			
est	quess for	Find Find	other sizes of the sizes of the sizes - Small - La	his image: arge e francaise agress	ee par un n	migrant	
est	guess for	Find Find this imag	other sizes of t zes - Small - La ge: <i>infirmier</i>	his image: arge re francaise agress	ee par un n	migrant	>
est	guess for	Find Fursh this imag	other sizes of t 2005 - Small - La ge: infirmier	his image: arge re francaise agress	ee par un n	migrant	>
est	guess for mme qui vers.france	Find Fursion this image frappe of 24.com/	ether sizes of the small - La small - La small - La small - La secondaria de la secondaria	his image: arge e francaise agress pres n'est pas un r ox-video-homme-frapp	ee par un n nigrant ma e-i ▼ Transi	nigrant tis un late this pag	
est ho	guess for mme qui vers.france2 0, 2017 - L'ho	Find this imag	other sizes of ti cos - Smail - La ge: <i>infirmien</i> ces infirmiè (20170320-into e de 38 ans, ag	his image: arge e francaise agress ress n'est pas un r ox-video-homme-frapp resse deux infirmières	ee par un n nigrant ma e-i → Transi dans un couloi	iis un late this pag ir de l'hôpital	ge de

https://www.francetvinfo.fr > Monde > Europe > Migrants - Translate this page

Mar 22, 2017 - Non, cet homme n'est pas un migrant qui frappe des infirmières dans un hôpital français. La vidéo, mise en ligne dans la soirée du 18 mars sur la page Facebook du groupe SOS Racisme-antiblanc, est devenue virale. Mais la vidéo a été postée avec un message qui divulgue une fausse information et ...

Visually similar images



Report images

G/A

Credit: Denis Teyssou (AFP)



Online Rumours Analysis





Online Rumours Analysis





PHEME Project



• Rumour is "a circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety"

Examples:

- True rumour: "10 people dead in Charlie Hebdo according to witnesses"
- False rumour: "GERMAN NEWS REPORT: Co-Pilot of Germanwings Airbus Was MUSLIM CONVERT ...'Hero of Islamic State'?"
- Unverified rumour: "Police in Ferguson claimed that Mike Brown had been involved in a robbery"



Rumour Stance Classification

(i)

• What is being said about a claim/post?

Tweet

Emmanuel Macron @EmmanuelMacron

Our house is burning. Literally. The Amazon rain forest - the lungs which produces 20% of our planet's oxyger - is on fire. It is an international crisis. Members of the G7 Summit, let's discuss this emergency first order in two days! #ActForTheAmazon



9:15 PM · Aug 22, 2019 · Twitter for iPhone

57.4K Retweets 159.9K Likes

Caneta Desesquerdizadora @Desesquerdizada

Replying to @EmmanuelMacron

FAKE NEWS. 1. This picture was taken by Loren McIntyre, an American photojournalist that DIED IN 2003. It's for sale here: alamy.com/stock-photo-am...

2. The real lungs of the World are algae. Up to 70% of Earth's oxygen is produced by them.



Stock Photo - Amazon rain forest afire Download this stock image: Amazon rain forest afire. - BNX7RP from Alamy's library of millions of high resolution stock photos, illustrations alamy.com

♡ 16.3K 10:52 PM - Aug 22, 2019
 ○ 3,380 people are talking about this
 deny
 ② ② ② ○ ③ ○ ③ ○ ③ ○ ② ○



Replying to @EmmanuelMacron

Mr Macron, the Amazon does not produce 20% of the planet's oxigen. Secondly, the photo you used is very old from years ago, that's called 'fake news' these days and does NOT help the Amazon. Last but not least, you have an approval rating of 22% in France, fix your house first.



>



Stance classification

• PHEME dataset

Dataset	Rumours	Supporting	Denying	Questioning	Commenting	Large
Ottawa shooting	58	161	76	64	481	\checkmark
Ferguson riots	46	192	83	94	685	\checkmark
Prince in Toronto	12	19	7	11	59	×
Charlie Hebdo	74	236	56	51	710	\checkmark
Ebola Essien	2	6	6	1	21	×
Germanwings crash	68	177	12	28	169	\checkmark
Putin missing	9	17	7	5	33	×
Sydney siege	71	89	4	99	713	\checkmark



Stance classification

- RumourEval 2019 shared task
 - Task A: stance classification (Comment/Support/Deny/Query)





https://competitions.codalab.org/competitions/19938



Disinformation Analysis Pipelines

Swedish Term Recognition Twitter (9) Veracity (1) Weish WeVerify 🗸



1,200 free requests / day Batch processing not available



Analysing Memes and Adverts: Example Election Ad

Inactive 60 ... 4 Dec 2019 - 8 Dec 2019 ID: 446445396070238 Conservatives Sponsored . Paid for by The Conservative Party We're investing more in our NHS so we can deliver worldclass care when you need it. VOTE CONSERVATIVE X 620 ads use this creative and text 0

https://www.facebook.com/ads/library/?active_status=all&a d_type=political_and_issue_ads&country=GB&view_all_pa ge_id=8807334278

- Misleading claim made in the ad image only
- 800k-900k impressions
- Misleading
 - it's <u>six</u> in the next parliament, and none is a completely new hospital
- <u>https://fullfact.org/election-2019/ads/</u>



Analysing Memes and Adverts: Example Election Ad (2)

Output

40 new hospitals nhs vote conservative

Source

Uploaded file: 40-new-hospitals-ad-misleading.jpg

Original image:



https://www.facebook.com/ads/library/?active_status=all&a d_type=political_and_issue_ads&country=GB&view_all_pa ge_id=8807334278

http://demos.gate.ac.uk/weverify/ocr/

- Misleading claim made in the ad image only
- 800k-900k impressions
- Misleading
 - it's <u>six</u> in the next parliament, and none is a completely new hospital
- <u>https://fullfact.org/election-2019/ads/</u>



Analysing Memes and Adverts: Try It

Go to <u>https://cloud.gate.ac.uk/shopfront/displayItem/ocr-service</u>

Paste this Image URL in the Test this pipeline field: <u>https://pbs.twimg.com/media/EtrCBYYXMAUwpl7?format=jpg&name=small</u>

Press the Test Pipeline button

Wait to see the results:

ocr_text cdc i over 500 deaths recorded follow i ng experimen tal mrna inj ections iza through january 29 7 2021

url <u>https://pbs.twimg.com/media/EtrCBYYXMAUwpl7?format=jpg&name=small</u>

Additional experiment: hunt the original image behind this meme by going to Google images and pasting the image URL in the search



Veracity classification

• The task

- $\circ\,$ Determine the veracity of an entire rumour
- Binary or three-class problem (true/false/unverified)
- Potentially less useful for professionals
- RumourEval 2019 shared task
 - Three-class: true/false/unverified

	Task A	
	Macro F	RMSE
System X	0.5765	0.6078
System Y	0.2856	0.7642
System Z	0.2620	0.8012





Let's try the veracity classifier via Google sheets

- The easiest way to try the services for smaller data samples
- For processing via GATE Cloud and for large volumes, see Module 3
- Currently in beta not yet a published add-on you can install to your own sheets
- For testing purposes take a copy of <u>this Google Sheet</u>
 - You can only *view* the original, not edit it
 - \circ go to File \rightarrow "Make a copy"



You may need authorisation

- In your copy of the sheet, "Add-ons" → "GATE Cloud Text Analysis" → "Open sidebar"
- First time you do this it will request authorisation
 - Your email will appear as the developer, this is normal!

uired	
iment needs your p	ermission to run.
Continue	Cancel
	uired Iment needs your p Continue



Google hasn't verified this app

The app is requesting access to sensitive info in your Google Account. Until the developer (<u>you@gmail.com</u>) verifies this app with Google, you shouldn't use it.







Authorisation

Continue only if you understand the risks and trust the developer (you@gmail.com).



 When the add-on is published properly then we will be the developer and it won't be considered "unsafe"

	Sign in – Google accounts	
VPN 🔒	accounts.google.com/signin/oauth/consent	⊳ ।
This	will allow GATE Cloud Text Analysis to:	
•	View and manage spreadsheets that this application has been installed	in (j
٥	Connect to an external service	i
	Display and run third-party web content in prompts and sidebars inside Google applications	(i)

Make sure that you trust GATE Cloud Text Analysis

You may be sharing sensitive info with this site or app. Find out how GATE Cloud Text Analysis will handle your data by reviewing its terms of service and privacy policies. You can always see or remove access in your **Google Account**.

Find out about the risks

Cancel





Using the add-on

- Once authorised, "open sidebar" again
- "Configure credentials" lets you supply your API key ID and password
- Intended workflow
 - Choose the column of texts to process
 - Typically, set destination immediately to the right of the top-most input cell
 - For input A2:A10, destination would be B2
 - Choose Rumour Veracity Classifier
 - Optional) Configure annotations you want to extract using column headings

Configuration

Configure credentials

Analyse Text

Range to analyse

Sheet1!A2:A10	Use selected range

Specify a rectangular range of cells containing the text to analyse

Destination

Sheet1!B2

Use current cell

Leading cell of the range where results should be placed

Process with service

Rumour veracity classifier

Description

User generated content such as tweets often make claims that are unsubstantiated and possibly untrue. This service attempts to classify whether a text is discussing a rumour that is true, false or

Full details (opens in new window)...

Configuration range

unverified.

Sheet1!B1:C1 Use selected range

Open configuration helper

Specify a rectangular range of cells (typically a set of row or column headers) specifying how to handle the results from the service.

Submit job





- The sheet has a few texts some true, some false
- Select the cells with input text (A2 to A8)
- Under "Range to analyse" click "use selected range"
- We will put the output alongside click into cell B2, then under destination click "use current cell"
 - Results from processing A2 will go in B2 & C2
 - A3 will go in B3 & C3
 - Etc.
- Select Rumour Veracity Classifier as the service to use
- Press Submit Job



Configuring annotations

- One service can produce many different annotations
- Approach taken here is to map output into several columns
- "Configuration range" (typically column headers) defines what to extract
- Scroll sidebar down, click
 "Open configuration helper"

		Use selected range
--	--	--------------------

Specify a rectangular range of cells (typically a set of row or column headers) specifying how to handle the results from the service.



Configuring annotations

• Click into cell B1, fill a horizontal range of 3 cells

Fill a	horizontal	range of	3	cells
G	30	Cancel		

- You will see B1 to D1 filled with headings, and the range selected - click "Use selected range"
 - Column headings are annotation types
 - O Veracity rumour_label (confidence as %) Veracity rumour_label Veracity confidence



Output on the sample document

Veracity rumour_label (confidence as %)	Veracity rumour_label	Veracity confidence
false (87.79%)	FALSE	0.8779187799
false (87.79%)	FALSE	0.877910912
true (82.08%)	TRUE	0.8207599521
unverified (94.59%)	unverified	0.945890367
unverified (97.00%)	unverified	0.9699695706
unverified (62.94%)	unverified	0.6293928027
unverified (57.41%)	unverified	0.5741012096

Domain Credibility





Anthony Fauci Admits Chances Of Getting Accurate Results From COVID PCR Tests Are 'Miniscule' bit.ly/2IyiAiI via @EUTimesNET



Anthony Fauci Admits Chances Of Getting Accurate Results From COVI... As reported by Jon Rappaport, during a July 16, 2020, "This Week in Virology" podcast, Anthony Fauci made a point of saying that the \dots \mathscr{S} eutimes.net

11:20 PM · Nov 13, 2020

🗇 1 🦻 See the latest COVID-19 information on Twitter

As an example, let's assume that we see this tweet and want to know how credible the claim it contains is.

The service checks the URL of the tweet and the URL within the tweet too.



(i)



After processing by the service we can see that....

DomainCredibility	
credibility-domain	eutimes.net
credibility-labels	hate,conspiracy
credibility-resolved-url	https://twitter.com/EUTimesNET/status/1327390787202654213
credibility-source	OpenSources
credibility-timestamp	20170428

The URL of the tweet has also been related to the eutimes.net domain and labelled as both hate and conspiracy by OpenSources. The short URL in the tweet is also resolved to eutimes.net and then flagged accordingly.



Domain Credibility



> We currently have aggregated data on:

- 1320 unique domains
- 724 domain to Twitter account mappings
- 918 domain to Facebook page/group mappings
- > Try it yourself at:

https://cloud.gate.ac.uk/shopfront/displayItem/domain-credibility





Domain Credibility in Sheets

- Let's add domain credibility information in column E
- Select the cells with input text (A2 to A8)
- Under "Range to analyse" click "use selected range"
- We will put the output alongside click into cell E2, then under destination click "use current cell"
 - Results from processing A2 will go in E2
 - A3 will go in E3
 - Etc.
- Select **Domain Credibility** as the service
- More steps on next slide



Domain Credibility in Sheets (2)

• Select E1 to configure the output range

Configuration range

Sheet1!E1

Use selected range

Open configuration helper

Specify a rectangular range of cells (typically a set of row or column headers) specifying how to handle the results from the service.

- Press Use Selected Range first (so it looks like above)
- Then press Open Configuration Helper, select as shown

E1 value should now be:

DomainCredibility credibility-domain = credibility-labels

Fill a	horizontal	\$	range of	1	< >	cells
(30	C	ancel			



Press Submit Job and get the output

E	
DomainCredibility credibility-domain = credibility-labels	*
infowars.com = conspiracy	
breitbart.com = political,unreliable,bias	



GATE General Architecture for Text Engineering **COVID-19 Misinformation Classifier**

Label Studio Labeling Annotation Guidelines Claim: An "open letter" purportedly written by US billionaire and philanthropist Bill Gates about the novel coronavirus pandemic has been shared in English and Chinese on Facebook, Twitter and various websites. The Chinese-language posts state the letter was translated from its original publication in British newspaper The Sun. Explanation: But the letter is fake; Gates' philanthropic organisation, the Bill & Melinda Gates Foundation, clarified on Weibo that the letter presents "false information" and urged netizens to stop sharing it. The Sun has reportedly removed the letter from its website. https://factcheck.afp.com/gates-foundation-urges-netizens-stop-sharing-fake-bill-gates-coronavirus-letter Choose one of the following (required): Public authority^[1] Community spread and impact^[2] Medical advice, self-treatments, and virus characteristics^[3] Prominent actors^[4] Further details: Conspiracies^[5] Virus transmission^[6] Virus origins^[7] https://arxiv.org/abs/2006.03354 Public preparedness^[8] Vaccines and medical treatments^[9] Protest and civil disobedience^[0] Social alarm and divisions[a] None of the above / Cannot determine^[q] Confidence (required): 0 1 2 3 4 5 6 7 8 9 Optional Remarks (optional, press enter to add remark)

Test the Classifier on GATE Cloud

Test this pipeline

Type the content to annotate:	The vaccine against the new coronavirus was created before the disease appeared; three people who recently arrived from China in Sarajevo Canton are under medical supervision and have brought COVID-19 to Sarajevo.
	 ©
Or select a text file:	Choose file No file chosen
Output type:	JSON
Document format:	plain text ~
Restore defaults	✓ MisinfoClass ✓ Attention □ Token
Test Pipeline	download
Annotation types: Attention	MisinfoClass
The vaccine against the new of have brought COVID-19 to Sa	coronavirus was created before the disease appeared; three people who recently arrived from China in Sarajevo Canton are under medical supervision and rajevo.
Annotations at this loc	ation Try it yourself:
MisinfoClass	https://cloud.gate.ac.uk/shopfront/displayItem/covid19-misinf
class Vacc	

WeVeri



COVID Categoriser in Sheets

- Let's add COVID-19 topic category information in column F
- Select the cells with input text (A2 to A8)
- Under "Range to analyse" click "use selected range"
- We will put the output alongside click into cell F2, then under destination click "use current cell"
 - \circ $\,$ Results from processing A2 will go in F2 $\,$
 - A3 will go in F3
 - Etc.

• Select COVID-19 Misinformation Categoriser as the service

- It works just as well categorising true COVID-19 claims
- More steps on next slide



Domain Credibility in Sheets (2)

• Select F1 to configure the output range

Configuration	range
---------------	-------

Sheet1IF1

Use selected range

Open configuration helper

Specify a rectangular range of cells (typically a set of row or column headers) specifying how to handle the results from the service.

- Press Use Selected Range first (so it looks like above)
- Then press Open Configuration Helper, select as shown
- F1 value should now be:

MisinfoClass class

Fill a	horizontal	\$	range of	1	<	cells
(Go	C	ancel			



Press Submit Job and get the output

MisinfoClass class	
GenMedAdv	
Vacc	- 6
Consp	
Vacc	
PubPrep	
PubPrep PromActs	

GA

E General Architecture for Text Engineering **Exploration and visualisation of COVID-19** misinfo

Conspiracies -				
Fact checked claim **	Factchecker ++	Topic ++	Engagement 1	
→ The coronavirus "snuck up on us," adding that it is "a very unforeseen thing."	PolitiFact	Conspiracies	32,312	
$\rightarrow~$ The Covid-19 was invented in 2004, as shown in a patent filed by the Institute Pasteur	20 Minutes	Conspiracies	30,960	
\rightarrow This is a phishing scam.	Fullfact.org	Conspiracies	25,771	
\rightarrow Trump's claims about how quickly WHO acted to announce humanto-human transmission	Washington Post Fact-Checker	Conspiracies	19,546	
→ American scientist Judy Mykowitz exposes facts about the US administration regarding coronavirus and accuses the US military of fabricating it in conjunction with the Wuhan Chinese laboratory.	Misbar	Conspiracies	18,682	
→ Claim that the coronavirus was created and patented in a French institute (Institut Pasteur).	AFP	Conspiracies	16,679	
→ South Korean government secretly requested for certain factory to produce one million facial masks to be sent to North Korea.	JTBC news	Conspiracies	12,517	
→ "The 'biological' lab in Wuhan where the COVID-19 virus was created was 'funded' by President Barak Hussein Obama in 2015 to the tune of \$3,800,000 American dollars.	FactCheck.org	Conspiracies	12,389	
→ "There was no real scientific basis for believing that" social distancing would be necessary, "since it had never been studied."	PolitiFact	Conspiracies	12,332	
→ Riots in the American streets, after the collapse of the American economy due to the Covid-19 pandemic.	Misbar	Conspiracies	12,192	

Joint work with First Draft:

https://datadeficits.firstdraftnews.org/

https://firstdraftnews.org/long-form-article/data-deficits/





- In columns G onwards add the configurations for TwitIE, i.e. Person, Organization, etc.
- Run it to obtain the NEs that appear in each of the examples in the respective columns
- Refer to module 3 on GATE Cloud if you cannot recall how to run TwittlE/ANNIE in the sheets
 - Their configuration and running is identical, just TwitIE is better on short texts



Further materials

- Home page: <u>https://cloud.gate.ac.uk</u>
- An overview of disinformation responses
 <u>https://en.unesco.org/publications/balanceact</u>
- COVID-19 disinfodemic: <u>https://en.unesco.org/covid19/disinfodemic</u>

Thank you!



GATE Cloud API Keys

- If you haven't already, <u>register now</u> for an account on GATE Cloud
 - If you have already registered, log in
- Go to "<your name> account" in the top right
- Scroll to the bottom and "Manage your API keys"
- Generate a new key

Ian Roberts's account | Log out | Help

API Keys

Manage your API keys for the GATE Cloud REST APIs.



GATE Cloud API Keys

- API key is in two parts, the "key ID" and the "password"
- Copy and paste both into a text editor for future reference
- Note the warning: if you lose the password you must generate a new key (changing both ID and password)

New API Key

These are the details of your newly generated API key. Make sure you have made a note of the password before leaving this page - **API key passwords cannot be recovered** and you will need to generate a replacement key if the password is lost.

Key ID	gcj7nx3h69lq	
Password	si6fegs5trxs8g	0ui1bg
Description	Generated at 202	1-02-14T18:47:
	Save changes	Return to the key management page