Module 10: Disinformation Analysis

Kalina Bontcheva (k.bontcheva@sheffield.ac.uk)
This session will be recorded

Recorded video will be available after this session
WeVerify project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825297

 TYPES OF INFORMATION DISORDER

**FALSENESS**

**Disinformation**
Fabricated or deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumours.

**Malinformation**
Deliberate publication of private information for personal or corporate rather than public interest, such as revenge porn. Deliberate change of context, date or time of genuine content.

**INTENT TO HARM**

**MISinformation**
Unintentional mistakes such as inaccurate photo captions, dates, statistics, translations, or when satire is taken seriously.

Very difficult to distinguish mainly between MIS- and DISinformation

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

1 - Creation
- Play on Social Divisions...
- Political beliefs
- Religious beliefs
- Social beliefs
- ... using Digital content
- Memes
- False images
- False discourses
- False videos
- Misleading content

2 - Propagation
- Artificial propagation
  - Automation (bots)
  - Semi-automation (trolls)
  - Tool-enabled propagation
  - Advertisement
  - Astroturfing
  - Genuine propagation

3 - Verification
- Detection
  - Monitoring
  - Automated detection
  - Verification
  - Content verification
  - Source verification

4 - Impact
- Dissemination
  - Impact Analysis
  - Responses design
  - Technological responses
  - Social responses
  - Legal responses

Source: STOA report 2019
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?
Examples

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

We Verify project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825297
The Trumpet of Amplification

Source: https://medium.com/1st-draft/5-lessons-for-reporting-in-an-age-of-disinformation-9d98f0441722
Misinformation in search results

Source: (Albright, 2018)
Same image, different disinformation narratives and languages

Credit: Denis Teyssou (AFP)
Google getting confused

Credit: Denis Teyssou (AFP)
Online Rumours Analysis

Shortly after 1am on Tuesday, August 9, @zadie posted a link to an image of the London Eye apparently alight with the heartfelt message: "Oh my God! This can't be happening! The tweet is quickly picked up.

Oh my god! This can't be happening at London Eye! #Londonriots #Londonriot #PrayforLondon http://twitpic.com/6372vo

@zadie, 18 followers
Tue 9 Aug 01:07
Online Rumours Analysis
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

- What is being said about a claim/post?
Stance classification

- PHEME dataset

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

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

<table>
<thead>
<tr>
<th>Task B</th>
<th>Macro F</th>
</tr>
</thead>
<tbody>
<tr>
<td>System A</td>
<td>0.6187</td>
</tr>
<tr>
<td>System B</td>
<td>0.6067</td>
</tr>
<tr>
<td>System X</td>
<td>0.5776</td>
</tr>
</tbody>
</table>

https://competitions.codalab.org/competitions/19938
Analysing Memes and Adverts:
Example Election Ad

https://www.facebook.com/ads/library/?active_status=all&ad_type=political_and_issue_ads&country=GB&view_all_page_id=8807334278

- Misleading claim made in the ad image only
- 800k-900k impressions
- Misleading
  - it’s six in the next parliament, and none is a completely new hospital
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:

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

https://www.facebook.com/ads/library/?active_status=all&ad_type=political_and_issue_ads&country=GB&view_all_page_id=8807334278

- Misleading claim made in the ad image only
- 800k-900k impressions
- Misleading
  - it’s six in the next parliament, and none is a completely new hospital
Go to https://cloud.gate.ac.uk/shopfront/displayItem/ocr-service

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

Press the Test Pipeline button

Wait to see the results:

ocr_text  cdc i over 500 deaths recorded following experimental mrna injections iza through january 29 7 2021

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

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
  ○ 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

<table>
<thead>
<tr>
<th></th>
<th>Macro F</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>System X</td>
<td>0.5765</td>
<td>0.6078</td>
</tr>
<tr>
<td>System Y</td>
<td>0.2856</td>
<td>0.7642</td>
</tr>
<tr>
<td>System Z</td>
<td>0.2620</td>
<td>0.8012</td>
</tr>
</tbody>
</table>
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 this Google Sheet
  - You can only view the original, not edit it
  - go to File → “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!

![Authorisation Required]

Google hasn’t verified this app

The app is requesting access to sensitive info in your Google Account. Until the developer ([you@gmail.com](mailto:you@gmail.com)) verifies this app with Google, you shouldn’t use it.
Authorisation

- When the add-on is published properly then we will be the developer and it won’t be considered “unsafe”
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
Demo

- 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”
Configuring annotations

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

  ![Fill a horizontal range of 3 cells](image)

- You will see B1 to D1 filled with headings, and the range selected - click “Use selected range”
  - Column headings are annotation types
    - Veracity rumour_label (confidence as %)
    - Veracity rumour_label
    - Veracity confidence
<table>
<thead>
<tr>
<th>Veracity rumour_label (confidence as %)</th>
<th>Veracity rumour_label</th>
<th>Veracity confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>false (87.79%)</td>
<td>FALSE</td>
<td>0.8779187799</td>
</tr>
<tr>
<td>false (87.79%)</td>
<td>FALSE</td>
<td>0.877910912</td>
</tr>
<tr>
<td>true (82.08%)</td>
<td>TRUE</td>
<td>0.8207599521</td>
</tr>
<tr>
<td>unverified (94.59%)</td>
<td>unverified</td>
<td>0.945890367</td>
</tr>
<tr>
<td>unverified (97.00%)</td>
<td>unverified</td>
<td>0.9699695706</td>
</tr>
<tr>
<td>unverified (62.94%)</td>
<td>unverified</td>
<td>0.6293928027</td>
</tr>
<tr>
<td>unverified (57.41%)</td>
<td>unverified</td>
<td>0.5741012096</td>
</tr>
</tbody>
</table>
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.
After processing by the service we can see that...

<table>
<thead>
<tr>
<th>DomainCredibility</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>credibility-domain</td>
<td>eutimes.net</td>
</tr>
<tr>
<td>credibility-labels</td>
<td>hate, conspiracy</td>
</tr>
<tr>
<td>credibility-resolved-url</td>
<td><a href="https://twitter.com/EUTimesNET/status/1327390787202654213">https://twitter.com/EUTimesNET/status/1327390787202654213</a></td>
</tr>
<tr>
<td>credibility-source</td>
<td>OpenSources</td>
</tr>
<tr>
<td>credibility-timestamp</td>
<td>20170428</td>
</tr>
</tbody>
</table>

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
- Press Use Selected Range first (so it looks like above)
- Then press Open Configuration Helper, select as shown
- E1 value should now be:

```cpp
domainCredibility credibility-domain = credibility-labels
```
Press Submit Job and get the output

<table>
<thead>
<tr>
<th>DomainCredibility</th>
<th>credibility-domain = credibility-labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>infowars.com</td>
<td>conspiracy</td>
</tr>
<tr>
<td>breitbart.com</td>
<td>political, unreliable, bias</td>
</tr>
</tbody>
</table>
COVID-19 Misinformation Classifier

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.


Further details:
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 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.

Annotations at this location

MisinfoClass

class  Vacc

Try it yourself:
https://cloud.gate.ac.uk/shopfront/displayItem/covid19-misinfo
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”
  - 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

- 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
Press Submit Job and get the output
Exploration and visualisation of COVID-19 misinfo

<table>
<thead>
<tr>
<th>Fact checked claim</th>
<th>Factchecker</th>
<th>Topic</th>
<th>Engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td>The coronavirus &quot;snuck up on us,&quot; adding that it is &quot;a very unforeseen thing.&quot;</td>
<td>PolitiFact</td>
<td>Conspiracies</td>
<td>32,312</td>
</tr>
<tr>
<td>The Covid-19 was invented in 2004, as shown in a patent filed by the Institute Pasteur</td>
<td>20 Minutes</td>
<td>Conspiracies</td>
<td>30,960</td>
</tr>
<tr>
<td>This is a phishing scam.</td>
<td>Fullfact.org</td>
<td>Conspiracies</td>
<td>25,771</td>
</tr>
<tr>
<td>Trump’s claims about how quickly WHO acted to announce human-to-human transmission</td>
<td>Washington Post Fact-Checker</td>
<td>Conspiracies</td>
<td>19,546</td>
</tr>
<tr>
<td>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.</td>
<td>Misbar</td>
<td>Conspiracies</td>
<td>18,682</td>
</tr>
<tr>
<td>Claim that the coronavirus was created and patented in a French institute (Institut Pasteur).</td>
<td>AFP</td>
<td>Conspiracies</td>
<td>16,679</td>
</tr>
<tr>
<td>South Korean government secretly requested for certain factory to produce one million facial masks to be sent to North Korea.</td>
<td>JTBC news</td>
<td>Conspiracies</td>
<td>12,517</td>
</tr>
<tr>
<td>“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.”</td>
<td>FactCheck.org</td>
<td>Conspiracies</td>
<td>12,389</td>
</tr>
<tr>
<td>“There was no real scientific basis for believing that” social distancing would be necessary, “since it had never been studied.”</td>
<td>PolitiFact</td>
<td>Conspiracies</td>
<td>12,332</td>
</tr>
<tr>
<td>Riots in the American streets, after the collapse of the American economy due to the Covid-19 pandemic.</td>
<td>Misbar</td>
<td>Conspiracies</td>
<td>12,792</td>
</tr>
</tbody>
</table>

Joint work with First Draft:

https://datadeficits.firstdraftnews.org/
https://firstdraftnews.org/long-form-article/data-deficits/
Extra hands on

- 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 TwittIE/ANNIE in the sheets
  - Their configuration and running is identical, just TwitIE is better on short texts
Further materials

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

Thank you!
GATE Cloud API Keys

- If you haven’t already, register now 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
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: si6fgs5trxs8g0ui1bg
Description: Generated at 2021-02-14T18:47:

[Save changes] [Return to the key management page]