541 Lab 1 - Disinformation and metrics

Required readings

We will cover most of these videos during lecture time (I have indicated which ones below), but I am including them all here so that you have everything in one place. In the lab, there might be questions on specifics from the required readings whereas the optional readings are more of a general help and for your own interest.

- **Lec 1** Renee DiResta "How do we know what's true anymore"
- **Lec 2 (partly)** Rachel Thomas "Disinformation" (32:51 - 1:26:00)
- PBS Facebook documentary (1 h video. Link if you're in the US)
- A framework for ethical decision making (15 min read, skip the entire section "2. Traditional arrangement of the field of ethics")
- Rachel Thomas "The Problem with Metrics" (10 min read. There is also a video on the same topic, but this is more similar to what we covered in class already)
- Guillaume Chaslot "How Algorithms Can Learn to Discredit 'the Media'"(5 min read. Chaslot formerly worked at YouTube and founder of algotransparency).

Optional readings

▶ Click to show

Optional deepfake readings

▶ Click to show

Submission instructions

rubric={mechanics:20}

You receive marks for submitting your lab correctly, please follow these instructions:

- Follow the general lab instructions.
- Click here to view a description of the rubrics used to grade the questions
- Push your `.ipynb` file to your GitHub repository for this lab (make at least three commits).
- Upload your `.ipynb` file to Gradescope.
- Include a clickable link to your GitHub repo for the lab just below this cell

    - It should look something like this https://github.ubc.ca/MDS-2022-23/DSCI_541_labX_yourcwl.
    - If you are working in a group, you can create you own (public) repo in the UBC-MDS organization and link that instead.
- All your written answers must be in your own words.
- You are not allowed to use generative AI tools to write your answers for you or simply paraphrase answer that you generate from these tools (that will lead to a failing grade), but you can use them to further understand the topics you are learning about.

https://github.com/gtmx23/20241101_dsci541_lab1.git

Overall writing quality

rubric={writing:20}

You will receive an overall writing grade for the entire lab instead of for each question. This is just a small part of your total grade, but please use the Jupyter Lab spell checker extension to catch typos and read through your text for grammatical errors before submitting (or paste it into Google Docs/MS Word/Grammarly. You don't need to type anything under this cell, it is just a placeholder to generate the grading rubric.

1. Short answer questions

Keep your replies brief, 1-3 sentences per question. Although these are short answer questions, don't copy answers from the readings, use your own words so that you practice learning these concepts. These will not be discussed during the lab.

Question 1.1

rubric={reasoning:60}

1. What is the difference between misinformation and disinformation?
2. Disinformation campaigns are not necessarily trying to push a particular political agenda. What is one general goal of these campaigns and how do they try to achieve it?
3. What is an online echo chamber?
4. What are deepfakes and GPT-2/3/4?
5. What is meant with the statement "If you can make it trend, you can make it true"?

6. What does Goodhart's law state? Give one example of a how this law might play out in a real life situation (could be a hypothetical or real example).

1. The difference between misinformation and disinformation is that misinformation is simply false information, and may not necessarily be targeted or have a negative intention, such as an intent to cause harm. On the other hand, disinformation is often targeted at another group, and is intended to mislead audiences.

2. While disinformation campaigns can take various forms, one general goal is to often cause confusion and tension in the general public, sometimes even a targeted population or community. This is often achieved by creating divisions by amplifying divisive issues, undermining trust by spreading misleading information or by trying to manipulate public opinion.

3. Online echo chamber refers to a situation in which individuals (or groups) engage in conversations and share perspectives that amplify their personal beliefs, while excluding marginalized or opposing viewpoints. This is often facilitated by social media, whose algorithm is intended to help engage with like-minded individuals.

4. Deepfakes are artificial media that is created using deep learning techniques, to manipulate or replace existing content. These medias, created by generative models are often so well created that it is almost impossible to distinguish a deepfake from a real media. Due to this reason, deepfakes are very harmful to exist on the free internet. GPTs are a series of language transformer models. Each version of GPT is trained to generate human-like text based on the input it receives.

5. This statement refers to the idea that in this era of social media and fact-paced, engaging content,the popularity of information can influence its perceived truthfulness. In other words, the sheer amount of likes, shares and comments can create an illusion of credibility, even if the information itself is misleading.

6. Goodhart's law states that: "When a measure becomes a target, it ceases to be a good measure". One example is where teachers begin chasing maximising test scores which is a measure which aims to measure students' learning, instead of actually focusing on improving the learning of students.

2. Discussion questions

This section asks you to expand a bit on your reasoning, but still aim to write succinct replies around one paragraph per sub-question. The goal of lab discussions are not to provide you with the right answers, but to help your discussion along. Your TA will assist in this by bringing up topics that you might not have thought of, ask questions to break the silence or

a dead end, and move the conversation along so that you have time to go through most questions. How useful the lab discussion is for your submission ultimately relies on that you actively contribute to the discussion and help your peers contribute and exchange ideas.

Some tips to make your discussions in lab more effective

It is easy to overlook the flaws of our own reasoning, so having a discussion with colleagues is an excellent opportunity to develop your thinking and receive feedback from someone who can provide an alternative perspective from your own. Nevertheless, many people don't know how to have an effective discussion, so I am sharing a few tips for you to be able to make the most out of this opportunity:

- Commit to learning, not "winning" debates.
- Comment in order to share information and develop arguments further, not to persuade.
- Listen respectfully, without interrupting, to try to understand each others' views.
  - Don't focus on what you are going to say next while someone else is talking.
- Challenge ideas, not individuals.
  - And be open to having your own ideas challenged.
- Think about as good arguments as possible against your position.
  - This is especially useful if many of your peers have the same opinion, help your group find angles that you might otherwise be missing.
- Allow everyone the chance to speak.
  - Politely ask members of your group about their opinion.
- Avoid assumptions about any member of the class or generalizations about social groups.
  - Be careful about asking individuals to speak on the behalf of their (perceived) social group.
- Be aware of logical fallacies, but avoid pointing them out in rude or disrespectful ways.

Question 2.1 -- Social media responsibilities

rubric={reasoning:100}

1. Are social media platforms with the ability to tailor the "front page/news feed" for each user (e.g. Facebook, Twitter, and YouTube) responsible for the content that they recommend? Should they have the same editorial duties as a news network or are they more like a kiosk selling newspapers?
2. Is it wrong by the social media platform to label some stories as "fake news" and some stories as real news? Why should they have the right to perform this

censorship? Isn't the best remedy to fake stories to let them be heard and discussed so that everyone can see that they are fake?

3. Write down two or three arguments against your stance on either bullet point 1 or 2 above. Try to make them as strong as possible. Can you understand how some might value these argument higher than the ones you wrote down or do you think this is a clear cut issue without any strong argument against your position?

1. Whether social media platforms bear responsibility for the content displayed to their users is an intricate issue. These platforms, merely at the core, serve as spaces for users to share their content. However, a critical concern arises from the recommendation algorithms employed by these platforms, as they can inadvertently create filter bubbles that reinforce existing beliefs and opinions, neglecting alternative perspectives. Thus, social media platforms do carry a certain degree of accountability. Unlike news networks, they neither curate news themselves nor do they face accountability in the same manner, yet they are more than mere kiosks selling newspapers also because of their recommendation systems. Hence, it is their responsibility to strike the right balance involving preserving users' opinions and freedom of speech while simultaneously monitoring, tagging, or removing misinformation propagated through social media.

2. It is not inherently wrong for platforms to classify stories spreading misinformation as "fake news". There exists a responsibility to combat the rapid increase of misinformation, especially when a story is undeniably false. However, when faced with the nuances of subjective opinions or ambiguous areas, social media platforms could adopt a more transparent approach. Instead of outright censorship, these platforms could choose to leave such stories open for discussion while providing links to related sources, allowing users to form informed opinions. For example the legalization of euthanasia in certain countries is a controversial topic and both sides of the argument should be available to the users as there is no clear cut black and white answer in this case. The most effective remedy for fake stories, perhaps, lies in a balanced approach. Platforms can tag them as fake, issue warnings, but retain the content, fostering an environment where users are not shielded from information but are empowered to critically engage with it.

3. In bullet point 2, we said that social media websites should label content spreading misinformation as “fake news” after fact checking . But here are the challenges to labeling or tagging fake news:

    - **Amount or Volume of Content Generated:** Platforms like Facebook have a daily active users of about 2.08 billion. At the core moderation takes effort and it involves costs. The amount of content generated on these social media simply exceeds the amount they monitor or moderate.

- **Context Dependency:** Interpreting content always has certain context attached to it. Parodies or satirical content can be mistaken for genuine information or mislabeled by automated tagging systems as they might not capture these nuances or details.

- **Challenge due to different Content Formats:** The content on social media is in the form of text, images and videos and each of the formats requires different algorithms for analysis of misinformation and that adds to the challenge of making a unified-fact checking system.

As such, it is not a clear cut issue.

Question 2.2 -- Metric based engagement

rubric={reasoning:100}

1. Do you think it is unethical to build an online platform using engagement as the main metric? Don't we want our platforms to be engaging?
2. What are some strategies for using metrics more responsibly and effectively?
3. Whose responsibility is it that online platforms use better metrics? You as the data scientist, the company leadership, the governmental branch that regulates these industries, some third party organization, or the consumer who should be able to control how they spend their time and what content they share?

1. Using engagement as the main metric to build an online platforms is not inherently unethical, but could be problematic. The engagement metrics are important for the platform to understand the users' interest so that the platform can improve the users experience, which is good for platform since more consumers coming in and potentially increases the revenue a platform might generate. However, if we only look at improving engagement without considering the quality of the content, this may lead to the spread of the explicit content, disinformation or harmful content. Moreover, an overemphasis on maximising engagement might have unintended negative consequences on users, such as addictive behaviors and mental health issues caused by spending too much time interacting with excessive and perhaps unhealthy content on the platform.

2. We think that perhaps including more diverse metrics could help. For instance, other than engagement, we propose the use of a 5 degree metric system to measure the content quality like user satisfaction or misinformation rates. Though these might not be straightforward to measure, this more diverse focus might help to improve the quality content on the platform and mitigate the negative impacts of content driven by maximising engagement metrics. In addition, instead of looking at the immediate

engagement, we can evaluate how the content affects users over time. Moreover, the platform can be more transparent and allow the users to understand why they are seeing certain content. Metrics can also be designed to assess the impact of content on users' mental health.

3. Data scientists, the company leadership, governmental branches that regulate these industries, third party organizations, and consumers of such platforms should all be responsible for a better use of metrics. Data scientists have a responsibility to develop and propose metrics that are ethically sound. Company leadership should prioritize responsible metrics that balance business goals with user well-being. This focus from the top will also make it easier for data scientists to implement better metrics, as the direction provided by the leadership should guide the data scientists in the company. Governments can play a role in setting standards and guidelines to ensure that online platforms operate responsibly. Third-Party Organizations can also audit online platforms to check on their ethical policies and the implementation of their policies, like the metrics they use. They can also make comparisons with baseline models to see if models used by companies are drifting without biases. Consumers should also be informed consumers, who are aware of how engagement metrics influence their experience and content, and use this information to regulate their own online behaviours.

Question 2.3 -- Individual responsibilities

rubric={reasoning:100}

1. Do you think it is OK to share an article after just reading the headline but not the entire content? Does it matter what topic the article is about?
2. Describe a few technological and personal solutions that you think would help reduce spread of disinformation?
3. In the questions above we mostly discussed how social media company's could stop spreading disinformation, but what about your responsibility as a consumer of this information? Shouldn't you be able to select what are credible sources and avoid being tricked into clicking and sharing disinformation?

1. No, it is not alright to do so. Particularly in the current media landscape, where clickbait is rife and consumers are incentivised to scroll quickly past articles and posts which do not immediately stand out to them, the primary purpose of headlines is arguably to grab attention, whatever the cost. As such, producers of online articles are also incentivised to make headlines which invite engagement but might be misleading, and the contents of an article might not always match the headline. Consumers like us thus

need to be careful and should read articles in full before sharing them. We feel that it does not matter what the topic of the article is, but that it is particularly important to be careful when sharing political articles since fake news tends to thrive in this area, since it often incites very emotional and potentially harmful reactions, and might also have severe impacts on global events, such as elections.

2. We have a few suggestions, as follows:

Personal Solutions

- As discussed above, consumers should be aware of potential signs of fake news to look out for, and should be mindful of their online behaviours, like deciding on what to share.
- Consumers should carry out fact checking before sharing or engaging meaningfully with articles.
- Consumers should check the sources of articles and news before sharing them. Creators should also cite their sources.
- Awareness programs to share about digital literacy skills should be set up, and individuals should also play a role in educating their friends and family members who might not have these skills, like older family members who are not used to using online platforms.

Technological Solutions

Our group has thought of a few solutions which utilise technology to combat the propagation of disinformation. In all these solutions, we would like to emphasise the importance of human involvement - we believe that even when technology is used, humans should always be involved in the implementation of these solutions, as we feel that at this point, humans are ultimately better able to screen for disinformation than machines, since they are aware of cultural and political contexts which purely technological tools might miss out on in their screening. This also adds transparency and explanability to decisions made by online platforms, which we feel is desirable and more fair. We do recognise that humans are also fallible, since humans have political biases which might cause them to flag some types of news as fake news more than other types of news, but we feel that at this point, humans are still the best solution. We would also like to propose that perhaps a group of diverse human moderators could be used to hopefully average out biases and mitigate this. 3. We as individual consumers should take responsibility for their part in propagating fake news, as consumers play a critical part in driving engagement. Their behaviour in online spaces, such as sharing articles, are ultimately the proxies used by online platforms to determine which types of content to promote or restrict, and irresponsible behaviour like sharing articles before reading their contents will skew the online experience for not only the consumers themselves, but also for others who use the platform. This might lead to the spreading of both disinformation and misinformation, not only to their own followers or friends, but with the effects also potentially rippling down to a larger group of users, as the shared article might be interpreted to be 'good' by the algorithm and shared to others outside the

individual's social circle. While it is hard to always identify disinformation accurately since disinformation campaigns have grown in sophistication with the advent of large language models and more sophisticated AI tools, minimal measures like reading articles in full prior to sharing them, fact checking articles and verifying sources, should always be exercised by individual consumers who do have the tools to identify credible sources and play a role in stopping the sharing of disinformation.

- A sentiment analyser could be used to screen if articles or other forms of content contain or aim to incite extreme emotions, which would have negative effects on consumers and are a flag for potential disinformation. Content which does not pass this initial screening should be put on hold and passed to human moderators, who check
- Likes and dislikes can be one form users can use to indicate whether an article/form of content is 'good' or not, and might be an indicator of whether the content incites emotional responses, which is common in posts used in disinformation campaigns. Tools can hence be used to flag content which is heavily liked or disliked, which should raise suspicion, and these forms of content could be temporarily restricted. Human moderators should check these articles and other forms of content before choosing to delete them or lift the restrictions.
    - On this note, taking inspiration from Kaggle's implementation of tracking upvotes, other metrics like the rate and consistency of likes/dislikes could be tracked, as sudden bursts of likes or dislikes might indicate collusion, or concerted efforts to boost or dampen the spread of articles, which might indicate that a disinformation campaign is under way.
- Information on the likes and dislikes certain posts/articles/other forms of content receive should be made transparent to consumers, such as by having an optional panel consumers can open to view more information on the engagement posts have received. This includes not only the number of likes and dislikes but also other information about the engagement posts receive, particularly the ratio of likes to dislikes and the time taken to like/dislike a video. This is as very quick likes and dislikes are suspicious, and this additional information might help consumers to deduce whether the likes/dislikes were organic or not, and thus make a decision on what they think of the quality of information in the content.
- Companies should punish people who share disinformation, such as by banning accounts which spread disinformation.
- Much like how YouTube added links to official sources and information on content creators (eg: indicating that the content was made by a doctor licensed in the US on a video) when the site sensed that creators were talking about the COVID-19 pandemic, or how 'verified' badges were added to verified celebrities on Twitter, technology could be used to help consumers to identify more credible and less credible sources, which should help in stopping the spread of disinformation.

One caveat to this is that we cannot expect everyone to be media literate, since not everyone has access to resources on digital or media literacy. Thus, it may be argued that

this standard of responsibility should not be applied to all consumers utilising online platforms. However, instead of weakening the responsibility of the individual for exercising caution, we feel that this strengthens the need for consumers who are indeed digitally literate, and thus should know better, to be even more responsible in being on guard against fake news, such as by reading articles in full before sharing them. In addition, given the reality of the prevalence of misinformation and disinformation online, individuals, governments, schools, and other third party organisations should also be more active about educating the public about media and digital literacy, such that more consumers are able to identify disinformation and stop its spread, and be 'immunised' to the negative influence of disinfluence.

Question 2.4 (Challenging)

rubric={reasoning:20}

Read this article on the use of metrics to promote content at Facebook.

1. Summarize the article in 3-4 sentences to demonstrate your understanding of their main findings.
2. Write a short paragraph about what you think about the analysis and how it was conducted. Does it seem like a reasonable way of measuring what the authors claim that they are measuring?
3. When investigating issues (ethical or otherwise), it is important that we go beyond what is immediately presented to us in a single article. Find out more about the citizen panel. Is there anything there that changes your interpretation of the results? Why/Why not?

More info about the report here if you are interested, but it is not required to answer the question.

1. The article discusses an investigation into Facebook's Widely Viewed Content Report. By using data collected using Citizen Browser, the authors compared it with Facebook's report, revealing discrepencies in the ranking of the content which was most widely viewed, largely due to the metrics Facebook used to assess how widely viewed a piece of content was. The authors found that Facebook ignored information about total number of impressions domains received over time, instead only using a metric which measured the unique users reached by a form of content, and not including a metric capturing the number of times a form of content was viewed overall, which impacted the ranking of the content. In fact, after accounting for this additional, more comprehensive measure, it was found that sensationalist articles were shown to users more often than mainstream articles. A more thorough documenting of Facebook's transparency reports was recommended to address the discrepencies.

2. We think the analysis done by author was a thoughtful and comprehensive examination of Facebook's content recommendation strategy. Though the investigation was limited by the small sample size compared to Facebook's sample size, the authors' use of different metrics like unique users, impressions, frequency created a more comprehensive and realistic view on what kinds of content appear on the feeds of average Facebook users, compared to the metrics used in Facebook's report, which were more limited and might not truly reflect the average user experience. However it is important to note that the data collected might not exactly represent the behaviour of actual Facebook users. The statistical measures, data visualisation, and correlation analysis used by the authors were also appropriately used and the conclusions drawn were sound. Overall we found the way of measuring the popularity of content the authors proposed to be reasonable, and convincingly argued.

3. The second article focuses on how Citizen Project is designed and how they used the data to analyse on the content which is shown to users. It provided details about how they collected data, who are the different age categories involved in this data and much more. They are few limitations on their project, by seeing the data we can see their are biased and its difficult to identify the backgrounds of people. These may influence the types of content displayed and interacted with by panelists on Facebook which can impact the project's finding on popularity of content. Despite of few limitations, I think the project provides valuable insights on content choices. Even though the sample has some bias and might not be completely representative of the general Facebook-using population, the interpretation of results does not completely change. Overall, the project helps us to find useful information on how the content is shown on facebook and its popularity.

Help us improve the labs

The MDS program is continually looking to improve our courses, including lab questions and content. The following optional questions will not affect your grade in any way nor will they be used for anything other than program improvement:

1. Approximately how many hours did you spend working or thinking about this assignment (including lab time)?

Ans:

2. Were there any questions that you particularly liked or disliked?

Ans: [Questions you liked]

Ans: [Questions you disliked]