

Module Four - Topic 2

Identifying and countering hate speech online

Duration	Approximately 30-45 minutes
Subjects	Social Science, Computer Science, Civic Education, Ethics
Summary	This topic focuses on communication through social media with attention towards the mediating role of AI technologies. It addresses the issue of identifying and countering hate speech online. It aims to support students in their daily use of social media by helping them to identify harmful practices/messages and to develop healthy and responsible online behaviours.

Learning objectives

Learners will be trained in the identification of intolerant/violent arguments and communication (hate speech) and of their relationship with negative stereotypes and forms of xenophobia and hostility towards specific targets, including ethnic and religious minorities. In this regard, learners will be supported in identifying some constants in the behaviors of digital haters, in the main targets they choose (including religious minorities), in the strategies adopted for the dissemination of hate speech online and on the spread through social media such as Twitter, Facebook, YouTube etc.

Topic 2 also aims at the development of argumentative skills for respectful dialogue from different perspectives on complex issues of social life or current controversial topics and for the deconstruction of hate speech through *counter-narratives*.

Activity overview

Activities	Modality	Level
Activity 4.2.1 Identifying online hate speech	In class / Online / at home	Basic

Activity 4.2.1 - Identifying and analysing online hate speech

Estimated duration	30-45 minutes (in class/online) + time at home
Requirements	None

Hardware and Software tools used	In class: - Pen and paper - Mentimeter (optional) - Computer for the teacher (optional) Online: - Online meeting platform - Shared document (e.g. Google Docs) - Mentimeter (optional) - Digital device for all
Type of assessment	Oral presentation
Glossary	Hate speech

Procedure

1. **Anticipatory Set:** The lesson starts with an introduction to key terms and notions regarding respectful online communication vs hate speech; cyberbullying; stereotypes, xenophobia, islamophobia, antisemitism online etc. The teacher shares examples of hate speech and/or instances where online hate speech has sparked debate or had consequences in the real world (see resources for teachers). The teacher also introduces students to the analysis of hate speech, identifying in the used examples some main elements: hater(s), targets, other roles involved (e.g. silent witnesses), content of the message, time and context, possible escalation of hatred, effect and reactions, etc. A guiding scheme may be the following:



2. **Independent Practice:** Students are asked to identify examples of hate speech and ways to fact check them - i.e. to assess whether they are hate speech or not. Every

student prepares a short written description of the example, with a link to an online resource that can be accessed in class during the following discussion.

3. **Guided practice and understanding:** The teacher provides an example of how to fulfil the task and how to present the example to the class. Students are also encouraged to focus on instances of hate speech they have encountered in their daily use of social media.
4. **Check for Understanding:** the class gets back together so that each student (or a restricted number of students, based on time availability) can present their example of hate speech online. After the presentations, the teacher moderates a discussion around how to identify and react to hate speech online.
5. **Closing:** The discussion can be closed with a conclusion, e.g. a shared definition of what online hate speech is and some recommendations to add to the code of conduct regarding how to identify fake news and how to respond to them (reporting mechanisms, not sharing, etc.)

Troubleshooting

Possible issue	Possible solution
Students take too much time to find materials / information overload	Students can be asked to focus on specific types of hate speech (e.g. cyberbullying) in order to avoid information overload during their research.
The discussion could be very chaotic or very silent	The teacher has an important role as the discussion moderator.
Suggestions on how to identify and respond to hate speech (code of conduct) could take up too much time	Part of this activity (code of conduct) could also be moved to Topic 4.4

Resources for teachers

Digital wellbeing

- <https://www.benesseredigitale.eu/pubblicazioni/>
- <https://www.mulino.it/isbn/9788815280992>
- Gui M, Büchi M. From Use to Overuse: Digital Inequality in the Age of Communication Abundance. *Social Science Computer Review*. 2021;39(1):3-19. doi:[10.1177/0894439319851163](https://doi.org/10.1177/0894439319851163)

Impact of online communication

- The impact of AI on communication networks and services, International Telecommunication Union,
https://www.itu.int/dms_pub/itu-s/opb/journal/S-JOURNAL-ICTF.VOL1-2018-1-PDF-E.pdf
- The Future of Human Communication: How Artificial Intelligence Will Transform the Way We Communicate,
<https://www.quantified.ai/blog/artificial-intelligence-in-communication>

Impact of social media on young adults:

- Moberg, M., & Sjö, S. (Eds.). (2020). Digital Media, Young Adults, and Religion: An International Perspective (1st ed.). Routledge. <https://doi.org/10.4324/9781351010610>
- Two social lives: How differences between online and offline interaction influence social outcomes,
https://escholarship.org/content/qt94n9w8b9/qt94n9w8b9_noSplash_293949a5e051fffc8e1fdcc9ffc168c4.pdf?t=qdtezbu
- Communicating across Cultures in Cyberspace, A bibliographical review of intercultural communication online, <https://core.ac.uk/download/pdf/12174297.pdf> \
- Some data and findings on Online Harassment
<https://www.pewresearch.org/internet/2014/10/22/online-harassment/>
- Council of Europe, “Model of Governance of Online Hate speech, On the emergence of collaborative governance and the challenges of giving redress to targets of online hate speech within a human rights framework in Europe ”,
<https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d>
- Resource from Council of Europe
<https://www.coe.int/en/web/freedom-expression/hate-speech>
- Council of Europe - tool for analysing hate speech:
<https://rm.coe.int/advanced-guide-toolkit-how-to-analyse-hate-speech/1680a217cd>
- Resource by Council of Europe on media’s contribution to social cohesion, intercultural dialogue, understanding, tolerance and democratic participation "Living Together".
<https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680483533>

MEMES as a “case study” of on-line hateful communication among youngsters

- Facebook data set and policy on scrutinizing memes post in FB groups against hate speech with the help of AI and algorithms,
<https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/>
- On the difficulties of AI to fully recognize hateful memes
<https://venturebeat.com/2020/12/01/ai-still-struggles-to-recognize-hateful-memes-but-its-slowly-improving/>
- “The hateful memes challenge: Detecting Hate speech in multimodal memes”,
<https://proceedings.neurips.cc/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf>
- “Detecting hateful memes using AI”,
<https://www.eng.ed.ac.uk/about/news/20201201/detecting-hateful-memes-using-ai>
- <https://www.drivendata.org/competitions/64/hateful-memes/>

Tools to counter cyberbullying

- Facebull (pen and pencil facebook showing social dynamics in social media such as cyberbullying)-

<https://educationaltoolsportal.eu/educationaltoolsportal/en/tools/facebull>

Country-specific resources on hate speech:

Greece

- “Detecting hate speech in Greece” <https://ecpr.eu/Events/Event/PaperDetails/59158>
- <https://www.loc.gov/law/foreign-news/article/greece-new-law-criminalizes-denial-of-genocide-hate-speech-and-other-acts-of-racism/>
- <https://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=15946&LangID=E>

Bulgaria

- “Mapping out the National Context of hate speech in Bulgaria”, <https://ec.europa.eu/migrant-integration/librarydoc/mapping-out-the-national-context-of-online-hate-speech-in-bulgaria> report in pdf format (<https://ec.europa.eu/migrant-integration/?action=media.download&uuid=E0C538B8-F34-6E1E-63F2DC364C4A5BF6>)
- <https://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=25881&LangID=E>
- Internet Hate speech – case study of Poland, Romania, India and Bulgaria <https://ieeexplore.ieee.org/document/9223209>

Romania

- “The Law of the Jungle. The on-line hate speech against the Roma in Romania”, <https://www.tandfonline.com/doi/pdf/10.1080/15564886.2021.1895391?needAccess=true>
- <https://balkaninsight.com/2021/03/31/racism-police-brutality-and-online-hate-why-romania-roma-are-no-nearer-their-black-lives-matter-moment/>

Italy

- <https://www.odiareticosta.it/>
- “Contro L’Odio”: A Platform for Detecting, Monitoring and Visualizing Hate Speech against Immigrants in Italian Social Media, <https://journals.openedition.org/ijcol/659>
- Hate Speech Annotation: Analysis of an Italian Twitter Corpus, <https://iris.unito.it/retrieve/handle/2318/1659197/387293/paper024.pdf>
- Amnesty Manuale su hate speech - conoscerlo e contrastarlo <https://d21zrvtkxtd6ae.cloudfront.net/public/uploads/2021/02/Amnesty-Manuale-hate-speech-2020-con-logo-1.pdf>