U6: KAMoT - Using tools to spot cyberbullying

Educational Toolkit

OVERVIEW

Areas covered:

Understanding	Preventing	Responding

Areas of SEL:

Self-awareness	Self-management	Social awareness	Relationship skills	Responsible decision-making

Audience:

11-13	14-18

Timing: **45 minutes**

Learning outcomes:

> **Learners will be able to:**
- **Identify variables/characteristics of cyberbullying behaviour.**
- **Consider the benefits and limitations of tools used to detect cyberbullying.**

Key vocabulary: **cyberbullying, hate speech, variables, patterns, algorithms, artificial intelligence (A.I.), monitoring, detecting, privacy, moderating.**

Resources: **Google Slides, copies of 'Cyberbully detector' worksheet (slide 6)**

Key questions:
- How can you spot a cyberbully or cyberbullying behaviour on social media?
- What are the characteristics of cyberbullying behaviour?
- Can cyberbullying be accurately detected in…
  - ...public online spaces? Why/why not?
  - ...private online spaces? Why/why not?
- What tools/methods could be used to detect cyberbullying?
- Should we use automated tools to detect cyberbullying? Why/why not?
- What actions should tools/networks take once cyberbullying is detected?

ACTIVITY PLAN:

Starter activity (10 minutes)

How to spot a cyberbully

Explain to young people that this session is about how online tools on social media could be used to detect patterns in cyberbullying behaviour or users who may be cyberbullying others.

Group the young people into small groups of 3-4. Ask each group to write down on a piece of paper the different ways they could tell if someone was showing cyberbullying behaviour on a social media platform.

Ask them to sort their suggestions under the following headings:
- **Textual** (words/phrases/language used to bully)
- **Visual** (use of images, videos and other visual communication e.g. emoji)
- **Aural** (audio in video and sound recordings)
- **Behavioural** (actions that can be seen on social media e.g. 'block', 'dislike')

After five minutes, ask groups to share some of their ideas and discuss where and how they saw/experienced these examples e.g. which social media platforms, who was displaying these behaviours, etc.
Ask young people which category had the most examples - it may be the case that 'textual' or 'visual' examples were the most common.
Ask young people if there are any types of cyberbullying behaviour that cannot be easily identified online (e.g. private/direct messages, vexatious reporting of other users, use of tools to block/mute someone, other forms of exclusion from a discussion or group.).

Activity (25 minutes)

Cyberbully detector

Ask young people if they are aware of any methods that social media platforms use to spot unacceptable behaviour (such as cyberbullying, hate speech, harassment, etc.). They may mention methods such as the use of moderators, automated tools/algorithms, keyword identification or machine learning (or artificial intelligence).

Explain that they are going to design their own 'Cyberbully detector'; an automated tool that will be used on a social media platform to automatically identify behaviour that could be cyberbullying.
Using the template on slide 6, ask young people to work in pairs to complete their plan. They should consider the following areas:
- Which social media platform the tool will be used on
- What types of bullying behaviour it will detect (e.g. textual/visual/aural/behavioural)
- How will the tool know/learn what to look for?
- How will the tool's work be checked for accuracy?

- Will the tool examine public content, private content, or both? (Remind young people about data protection laws - companies may only access private communications on their platform if users have explicitly and knowingly given consent!)

After developing their ideas, ask some pairs to share their 'Cyberbullying detector' and explain it in detail.

Ask all pairs to consider up to three strengths and three weaknesses of their tool, and record these on their worksheet.

Plenary (10 minutes)

Ask learners what they would like their tool to do after it has detected cyberbullying behaviour. Collect some suggestions of possible actions e.g.
- Issue a warning to the offensive user
- Alert a human moderator to review the content
- Delete or remove the content from being viewed by other users
- Send a message to the target offering help/support