TELUS Averti[MD]

Le défi de modération de l'IA

Les élèves jouent le rôle de « modérateurs de contenu IA » comme s'ils étaient les algorithmes des médias sociaux. Ils reçoivent des exemples de publications ou de commentaires (préparés par l'enseignant) et doivent décider si le contenu doit être signalé, retiré ou autorisé. Au cours de la discussion, ils explorent leurs différences de jugement et expliquent pourquoi il est si difficile pour l'IA (et les humains) d'interpréter le ton, le sarcasme et le contexte.

Pour les élèves de la 7e à la 12e année

Objectifs d'apprentissage :

À la fin de cette activité, les élèves seront en mesure de :

- Comprendre le fonctionnement des outils de modération de l'IA et leur rôle dans la sécurité en ligne.
- Reconnaître pourquoi les systèmes automatisés ont du mal avec les nuances, le sarcasme et le contexte culturel.
- Réfléchir à l'empathie, à l'équité et à la responsabilité humaine derrière la modération en ligne.
- Développer la littératie numérique et la pensée critique autour du contenu en ligne.

Matériel nécessaire :

- Des copies imprimées ou numériques des exemples de publications sur les médias sociaux.
- Fiche de décision du modérateur (Signaler / Retirer / Autoriser; prévoir de l'espace pour le motif).
- Tableau blanc ou projecteur pour la discussion :

  Facultatif: Outil de sondage ou test éclair en ligne

  Courte vidéo explicative sur la modération

Exemples de lien : L'IA générative peut-elle remplacer les modérateurs de contenu? **(en anglais)**

Instructions

Commencez la séance en expliquant que de nombreuses entreprises de médias sociaux utilisent des systèmes d'IA pour détecter les contenus nuisibles ou offensants. Expliquez que l'IA utilise des modèles et des mots-clés pour détecter les publications nuisibles, mais qu'elle ne peut pas interpréter le ton ou les émotions – les humains doivent quand même faire des choix éthiques. Ou utilisez une vidéo pour l'expliquer.

Dites aux élèves qu'aujourd'hui, ils seront l'IA - mais en utilisant leur jugement humain - pour voir comment leurs jugements diffèrent et comment cela entre en jeu.

1. Activité du modérateur (20 minutes)

Divisez les élèves en groupes de trois à quatre personnes. Fournissez à chaque groupe les exemples de messages ci-dessous. Pour chacun d'entre eux, ils doivent décider : Autoriser ✅, Signaler ⚠️ ou Retirer 🚫 et expliquer leur raisonnement. Chaque groupe doit travailler en équipe pour décider. S'il y a désaccord, ils doivent en discuter jusqu'à ce qu'ils puissent en venir à une solution.

Publication 1 : Commentaire sous une vidéo sur le compte Instagram de Joel.

Eli_Speaks Belle présentation, Joel 😅
tu avais l'air plutôt intelligent cette fois-ci!
il y a 2 heures

Publication 2 : Commentaire laissé sur une publication Instagram par Maya montrant sa nouvelle coupe.

Maya-Bestie Si vous voulez critiquer la nouvelle coupe de cheveux de Maya, dites-le -lui en face, pas derrière son dos 💬✋
il y a 3 heures

Publication 3 : Selfie de groupe dont un camarade de classe est clairement exclu.

FriendsOnly Seuls les vrais amis savent ce qui s'est passé vendredi soir 😉💬
il y a 3 heures

Publication 4 : Commentaire publié sur la publication d'un centre communautaire au sujet d'un récent concours d'art mettant en valeur le projet de Sam.

ArtFan_22 Ce concours d'art était clairement truqué. Sans vouloir offenser, mais le dessin de Sam n'était même pas si bon 😬
il y a 5 heures

Publication 5 : Image générée par l'IA montrant un enseignant comme un méchant de jeu vidéo, sous-titrée.

GameEdits Boss Niveau : M. K.
il y a 6 heures

Publication 6 : Publié dans une histoire de classe avec des émojis rieurs.

HallwayHighlights
POV: Le lundi te frappe comme 😵💥
il y a 1 heure

Publication 7 : Clip de style TikTok.

TrendAlert Il n'est pas clair s'ils chantent à propos de quelqu'un en particulier ou s'ils passent simplement un bon moment.
il y a 1 heure

Publication 8 : Commentaire laissé sur un clip TikTok d'un camarade de classe chantant une chanson d'Ed Sheeran..

SingItLoud Ta voix est un peu fausse mais pas terrible 😂 continue à t'entraîner et tu y arriveras!
il y a 5 heures

Publication 9 : Publication sur Threads.

ConcernedPeer Est-ce que quelqu'un sait si Ava va bien? J'ai entendu dire qu'il s'était passé quelque chose après l'école! 😟
il y a 2 heures

Publication 10 : Publication sur Threads.

LaughsOnly J'ai hâte de disparaître pour toujours 😬 lol je plaisante.
il y a 1 heure

Fiche de décision du modérateur

Consignez les décisions de votre équipe ci-dessous. Soyez prêt à expliquer votre raisonnement lors de la verbalisation.

Publication	Décision (✅ / ⚠️ / 🚫)	Motif de la décision
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

Questions de réflexion :

Après avoir relevé le défi, discutez ou écrivez des réponses à ces questions :

- Quelle publication vous a posé le plus de difficulté? Pourquoi? Votre groupe a-t-il été d'accord sur toutes les décisions? Quelle a été la source des désaccords?
- Qu'est-ce que l'empathie peut nous aider à remarquer qui pourrait échapper à l'IA?
- Comment les outils de modération de l'IA pourraient-ils être améliorés pour rendre les espaces en ligne plus sûrs?

2. Récapitulation de la discussion (15 minutes)

a. Rassemblez tous les élèves et discutez de chaque publication et de la décision de chaque groupe. Les élèves doivent être en mesure de discuter du raisonnement qu'ils ont écrit.

b. Posez les questions suivantes :

   i. Quels messages ont été les plus difficiles à juger?

   ii. Votre groupe était-il d'accord sur tout?

   iii. Un algorithme d'IA ferait-il les mêmes choix?

   iv. Quel rôle l'empathie joue-t-elle dans le choix de ce qu'il est acceptable de partager?

3. Élaborez vos propres règles de modération (5 à 10 minutes)

a. En tant que classe complète, les élèves travaillent ensemble pour définir quatre ou cinq règles qu'ils utiliseraient pour former un modérateur d'IA (ex., détecter les émojis méchants, bloquer les injures).

b. Utilisez un tableau blanc ou un projecteur pour cela afin que la classe puisse travailler ensemble. Les décisions de modération de contenu nécessitent un jugement humain.

Reflexion et conclusion

Les élèves doivent comprendre que, bien que les modérateurs d'IA puissent être utiles et efficaces, la plupart des décisions de modération de contenu nécessitent un jugement humain pour être pleinement précises. Dans leur utilisation des médias sociaux, ils devraient assurer leur propre modération : lire avec empathie, réfléchir avant de partager.