# Revue de la littérature : Détection du cyberharcèlement à l’aide de l’intelligence artificielle

Le cyberharcèlement est devenu un problème majeur dans l’ère numérique, affectant particulièrement les jeunes sur les réseaux sociaux. Face à cette menace croissante, la recherche s’est intensifiée pour développer des systèmes de détection automatique à l’aide de l’intelligence artificielle (IA). Cette revue de la littérature explore les modèles, les méthodologies et les défis liés à la détection du cyberharcèlement, en s’appuyant sur douze sources académiques récentes.

## 1. Définitions et enjeux du cyberharcèlement

Le cyberharcèlement se définit comme un comportement agressif, répété et intentionnel visant à nuire à autrui par le biais de moyens électroniques (Salawu et al., 2020). Ses conséquences psychologiques sont graves, allant de l’anxiété à la dépression, voire au suicide (Balakrishnan & Kaity, 2023). Les plateformes sociales telles que Facebook, Twitter et Instagram sont les principaux vecteurs de ce phénomène (Khafajeh, 2024).

## 2. Modèles d’apprentissage automatique et profond

Les premières approches reposaient sur des modèles d’apprentissage supervisé comme les SVM, Naïve Bayes et les arbres de décision (Salawu et al., 2020). Cependant, les modèles d’apprentissage profond tels que CNN, RNN, LSTM et BERT ont démontré une meilleure performance dans la classification des contenus harcelants (Hasan et al., 2023 ; Khafajeh, 2024). Par exemple, Muralidhar (2024) a montré que BERT atteint une précision de 94 % dans la détection de propos abusifs.

## 3. Jeux de données et prétraitement

Les jeux de données utilisés proviennent de plateformes comme Twitter, Facebook et Kaggle. Ils sont souvent déséquilibrés, avec une majorité de contenus non harcelants (Ogunleye & Dharmaraj, 2023). Le prétraitement inclut la suppression des URL, des emojis, la lemmatisation et la tokenisation (Hasan et al., 2023). Philipo et al. (2024) soulignent l’importance de disposer de jeux de données multilingues et annotés de manière fiable pour améliorer la généralisation des modèles.

## 4. Modèles avancés et architectures hybrides

Des architectures hybrides comme CNN-LSTM ou BERT-SVM ont été proposées pour combiner les avantages des différents modèles (Saini et al., 2023). RoBERTa, une version optimisée de BERT, a montré des performances supérieures dans plusieurs études (Dubey et al., 2024). L’intégration de l’explicabilité via SHAP et LIME permet de mieux comprendre les décisions des modèles (Prama et al., 2025).

## 5. Défis et perspectives

Les principaux défis incluent la détection de propos implicites, la prise en compte du contexte, la multimodalité (texte, image, vidéo) et les biais algorithmiques (Batani et al., 2022). La confidentialité des données et l’éthique de la surveillance automatisée sont également des préoccupations majeures (Philipo et al., 2024). Les recherches futures devraient viser à développer des systèmes robustes, explicables et adaptés aux environnements multilingues et multiculturels.

## Conclusion

La détection du cyberharcèlement par l’intelligence artificielle représente une avancée majeure pour la protection des utilisateurs en ligne. Les modèles d’apprentissage profond, notamment BERT et ses variantes, offrent des performances prometteuses. Cependant, des efforts supplémentaires sont nécessaires pour surmonter les défis liés aux données, à l’éthique et à la diversité linguistique.

## Références (style APA)

Balakrishnan, V., & Kaity, M. (2023). Cyberbullying detection and machine learning: A systematic literature review. Artificial Intelligence Review, 56, 1375–1416.

Batani, J., Mbunge, E., Muchemwa, B., et al. (2022). A Review of Deep Learning Models for Detecting Cyberbullying on Social Media Networks. Springer.

Dubey, S., Pawar, L., & Suryavanshi, P. (2024). Cyberbullying Detection using Robustly Optimized BERT Pre-training Approach. IJRI.

Hasan, M. T., Hossain, M. A. E., Mukta, M. S. H., et al. (2023). A Review on Deep-Learning-Based Cyberbullying Detection. Future Internet, 15(5), 179.

Khafajeh, H. (2024). Cyberbullying Detection in Social Networks Using Deep Learning. IAJIT.

Muralidhar, A. (2024). BERT-Based Detection of Cyberbullying in Online Texts. SCSA Journal.

Ogunleye, B., & Dharmaraj, B. (2023). The Use of a Large Language Model for Cyberbullying Detection. Analytics, 2, 694–707.

Philipo, A. G., Sarwatt, D. S., Ding, J., et al. (2024). Cyberbullying Detection: Exploring Datasets, Technologies, and Approaches. arXiv.

Prama, T. T., Amrin, J. F., Anwar, M. M., & Sarker, I. H. (2025). AI Enabled User-Specific Cyberbullying Severity Detection with Explainability. arXiv:2503.10650.

Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. IEEE Transactions on Affective Computing, 11(1), 3–24.

Saini, H., Mehra, H., Rani, R., et al. (2023). Enhancing cyberbullying detection: A comparative study of ensemble CNN–SVM and BERT models. Social Network Analysis and Mining, 14(1).

Sridevi, M., Chandrasekaran, S., et al. (2022). Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media. Computational Intelligence and Neuroscience, Article ID 2163458.