UQAC
Université du Québec
à Chicoutimi

SOSAFE : UNE APPROCHE MULTIMODALE POUR LA DÉTECTION DES CONTENUS INAPPROPRIÉS POUR LES ENFANTS SUR TIKTOK

PAR OUSMANE DIALLO

MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI EN VUE DE L'OBTENTION DU GRADE DE MAÎTRISE SCIENCES (M.SC.) EN INFORMATIQUE

QUÉBEC, CANADA

RÉSUMÉ

La modération automatique des contenus destinés aux enfants sur TikTok constitue un défi particulier en raison du format court des vidéos et de l'importance des dynamiques sociales. Les approches existantes, majoritairement conçues pour des plateformes comme YouTube, peinent à capturer des signaux nuisibles souvent implicites et distribués entre plusieurs modalités.

Ce travail propose SoSAFE, une approche multimodale de détection de contenus inappropriés sur TikTok, fondée sur l'enrichissement du dataset TikHarm par l'intégration des descriptions, des commentaires et des métadonnées d'engagement, en complément des signaux audiovisuels. Une architecture multimodale unifiée est introduite, combinant des encodeurs pré-entraînés et un module de fusion avancé reposant sur un espace latent commun, un mécanisme de gating par modalité, une cross-attention inter-modale parcimonieuse et une agrégation finale par BiLSTM.

Les résultats expérimentaux montrent que l'approche proposée surpasse les baselines de l'état de l'art, avec une accuracy de 87.13% et un F1-score macro de 87.14% sur TikHarm. L'étude d'ablation confirme l'apport des modalités sociales et la pertinence des choix architecturaux retenus. Ce travail met en évidence l'intérêt d'une modélisation multimodale contextualisée pour la modération automatique sur TikTok.

ABSTRACT

The automatic moderation of content intended for children on TikTok presents a particular challenge due to the short-video format and the importance of social dynamics. Existing approaches, mostly designed for platforms such as YouTube, struggle to capture harmful signals that are often implicit and distributed across multiple modalities.

This work proposes SoSAFE, a multimodal approach for detecting inappropriate content on TikTok, based on the enrichment of the TikHarm dataset through the integration of descriptions, comments, and engagement metadata, in addition to audiovisual signals. A unified multimodal architecture is introduced, combining pretrained encoders with an advanced fusion module based on a shared latent space, modality-specific gating mechanisms, sparse inter-modal cross-attention, and final aggregation through a BiLSTM.

Experimental results show that the proposed approach outperforms state-of-the-art baselines, achieving an accuracy of 87.13% and a macro F1-score of 87.14% on TikHarm. The ablation study confirms the contribution of social modalities and the relevance of the selected architectural choices. This work highlights the value of contextualized multimodal modeling for automatic content moderation on TikTok.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

CV Cross Validation
V Vidéo
A Audio
C Commentaires
M Métadonnées
D Description
AMP Précision mixte automatique (Automatic Mixed Precision)
AP Précision moyenne (Average Precision)
ASR Reconnaissance automatique de la parole (Automatic Speech Recognition)
BiLSTM Réseau LSTM bidirectionnel (Bidirectional Long Short-Term Memory)
CLIP Pré-entraînement contrastif image–texte (Contrastive Language–Image Pre-training)
EMA Moyenne mobile exponentielle (Exponential Moving Average)
F1 Score F1 (moyenne harmonique de la précision et du rappel)
E-MTN Enhanced Multimodal Transformer Network
MMOB Multimodal Malicious Or Benign
HICCAP Hierarchical Cross-modal Context-Aware Pooling
ffmpeg Outil multimédia de traitement audio et vidéo
GPU Processeur graphique (Graphics Processing Unit)
I3D Réseau convolutionnel 3D inflaté (Inflated 3D Convolutional Network)
LSTM Mémoire à long court terme (Long Short-Term Memory)
MCC Coefficient de corrélation de Matthews
MLP Perceptron multicouches (Multilayer Perceptron)
mBERT BERT multilingue (Multilingual Bidirectional Encoder Representations from Transformers)
OCR Reconnaissance optique de caractères (Optical Character Recognition)
RAM Mémoire vive (Random Access Memory)
GRU unité récurrente à portes (Gated Recurrent Unit)
ROC Caractéristique de fonctionnement du récepteur (Receiver Operating Characteristic)
AUC Aire sous la courbe (Area Under the Curve)
URL Uniform Resource Locator
ViViT Video Vision Transformer
VGG16 Visual Geometry Group 16 couches
API Interface de Programmation d'Applications

DÉDICACE

À mon père, pour sa confiance inébranlable, son soutien constant et cette force silencieuse qui m'a appris à tenir debout, même dans les moments de doute. Sa foi en moi a souvent été plus grande que la mienne, et c'est grâce à elle que j'ai continué à avancer.

À ma mère, pour sa patience, son courage et les valeurs qu'elle m'a transmises dès l'enfance. Elle m'a appris à me battre, à être indépendante et à croire en mes capacités, même lorsque le chemin semblait difficile. Son amour et ses enseignements ont profondément façonné la personne que je suis aujourd'hui.

À mes frères et sœurs, pour leur présence, leur soutien et leur affection. Ils ont été une source de force, d'équilibre et de motivation tout au long de ce parcours. Savoir qu'ils étaient là m'a donné le courage d'aller jusqu'au bout.

À mon tonton Mohamed Moussa, qui m'a accueillie et soutenue en terre étrangère. Sa bienveillance, son aide et sa présence ont été précieuses et ont rendu ce chemin loin des miens plus doux et plus supportable.

Ce travail est dédié à ma famille, qui a été mon socle, ma force et ma motivation, et sans laquelle rien de tout cela n'aurait été possible.

Qu'Allah vous récompense pour chaque sacrifice, vous accorde Sa protection et fasse de ce travail une source de bien pour vous comme pour moi. Amine.

REMERCIEMENTS

Je rends grâce à Allah, le Tout-Puissant, pour la force, la patience et la persévérance qu'Il m'a accordées tout au long de ce travail, et sans lesquelles l'aboutissement de ce mémoire n'aurait pas été possible.

Je tiens à exprimer ma profonde gratitude à mes directeurs de mémoire, la professeure **Haïfa Nakouri** et le professeur **Hamdi Ben Abdessalem**, pour leur encadrement rigoureux, leur disponibilité et la pertinence de leurs conseils. Leur accompagnement attentif, leurs orientations scientifiques et leurs remarques constructives ont grandement contribué à la qualité de ce travail et à l'enrichissement de ma réflexion.

Mes remerciements s'adressent également à **l'Université du Québec à Chicoutimi**, ainsi qu'au corps professoral du programme de maîtrise en informatique, pour la qualité de la formation dispensée et pour l'environnement académique et scientifique favorable qu'ils offrent aux étudiants.

Mes sincères remerciements vont aussi à toutes les personnes qui ont, de près ou de loin, contribué à la réalisation de ce mémoire, que ce soit par leur aide, leurs conseils, leurs encouragements ou leur soutien moral.

Enfin, je tiens à remercier sincèrement ma famille et mes amis pour leur soutien constant, leurs encouragements et leur compréhension tout au long de ce parcours académique.

AVANT-PROPOS

Ce mémoire est né d'une préoccupation personnelle face à l'exposition croissante des enfants aux contenus diffusés sur les plateformes de réseaux sociaux. À l'ère des vidéos courtes et des algorithmes de recommandation, les frontières entre divertissement, information et contenus potentiellement nuisibles deviennent de plus en plus floues, en particulier pour les publics les plus jeunes. TikTok, par son format attractif et sa popularité massive auprès des enfants et des adolescents, occupe aujourd'hui une place centrale dans les usages numériques. Si cette plateforme constitue un espace de créativité et d'expression, elle expose également les mineurs à des contenus qui peuvent porter atteinte à leur bien-être psychologique, émotionnel et social. La protection de l'enfance et la préservation de l'innocence dans ces environnements numériques représentent ainsi un enjeu majeur, tant sur le plan sociétal que scientifique.

Ce travail s'inscrit dans une volonté de contribuer, à mon échelle, à la construction d'un environnement numérique plus sûr pour les enfants. En m'intéressant aux limites actuelles des systèmes de modération automatique, j'ai cherché à explorer comment les approches multimodales en intelligence artificielle peuvent mieux appréhender la complexité des contenus diffusés sur TikTok, en tenant compte non seulement des signaux audiovisuels, mais également du contexte textuel et social. Au-delà de l'aspect technique, ce mémoire reflète une conviction personnelle : la technologie, lorsqu'elle est conçue de manière responsable, peut devenir un outil au service de la protection des plus vulnérables.

Ce travail se veut ainsi une contribution scientifique, mais également une démarche engagée en faveur d'une utilisation plus éthique et plus consciente des technologies numériques, au bénéfice des enfants et de leur droit à un espace d'expression sécurisé.

INTRODUCTION

0.1 CONTEXTE GÉNÉRAL

La croissance rapide des réseaux sociaux au cours de la dernière décennie a profondément transformé les modes de communication, de divertissement et d'accès à l'information. L'évolution vers des formats visuels et interactifs a favorisé l'émergence de plateformes centrées sur la vidéo courte, dont les systèmes de recommandation algorithmique structurent fortement l'exposition aux contenus et les dynamiques d'engagement (Anderson, 2020). Ces environnements numériques occupent désormais une place centrale dans les pratiques sociales des jeunes générations, constituant un espace privilégié de socialisation, d'expression et de consommation de contenus.

Dans ce contexte, TikTok s'est imposée comme l'une des plateformes les plus influentes à l'échelle mondiale, en particulier auprès des enfants et des adolescents. En juin 2025, l'application a enregistré plus de 37 millions de téléchargements et s'est classée parmi les applications les plus téléchargées au niveau mondial (Slotta, 2025). Son format immersif, combinant vidéos brèves, musique, effets visuels et interactions sociales (commentaires, mentions « j'aime », partages), favorise une consommation continue et personnalisée des contenus, amplifiée par des mécanismes de recommandation fondés sur l'apprentissage automatique.

Si cette dynamique soutient la créativité et la viralité des contenus, elle soulève également des enjeux majeurs en matière de protection des mineurs. Plusieurs travaux ont documenté la diffusion de tendances virales dangereuses et de défis à risque sur TikTok, mettant en évidence des conséquences sanitaires concrètes chez les populations pédiatriques (Kriegel *et al.*, 2021; Minhaj & Leonard, 2021). L'exposition à des contenus violents, sexualisés ou liés à des thématiques sensibles telles que l'automutilation et le suicide constitue ainsi une préoccupation importante dans les débats sur la régulation des plateformes numériques.

Au-delà de sa dimension technique, TikTok peut être appréhendée comme un système socio-technique caractérisé par une ambivalence structurelle entre créativité, engagement et exposition à des risques potentiels (López, 2021).

Face à ces enjeux, les plateformes ont mis en place des dispositifs de modération combinant règles communautaires, signalement par les utilisateurs et systèmes automatisés fondés sur l'intelligence artificielle. Toutefois, la détection fiable de contenus inappropriés demeure un défi scientifique et technique majeur. Le volume de données produites, la rapidité de diffusion des vidéos et leur nature intrinsèquement multimodale rendent l'analyse automatique particulièrement complexe.

Dans ce contexte, l'approche multimodale apparaît comme une voie particulièrement pertinente pour appréhender la complexité des contenus diffusés sur les plateformes de vidéos courtes. En intégrant conjointement des signaux visuels, audio, textuels et comportementaux, les méthodes d'apprentissage automatique peuvent exploiter des informations complémentaires et améliorer la robustesse des systèmes de détection. Cette perspective suggère que la nocivité d'un contenu sur TikTok est fondamentalement socio-contextuelle et ne peut être réduite à un signal isolé, mais doit être analysée à travers l'interaction structurée de plusieurs modalités. Le développement de telles approches constitue ainsi un enjeu scientifique et sociétal majeur, visant à contribuer à un environnement numérique plus sûr pour les jeunes utilisateurs. Malgré ces défis, les solutions actuelles de modération présentent encore des limites importantes.

0.2 MOTIVATION DE RECHERCHE

Malgré les progrès récents en apprentissage automatique, la modération automatique des contenus sur les plateformes de vidéos courtes, et en particulier sur TikTok, demeure un problème largement ouvert. Les travaux existants se concentrent principalement soit sur

des approches unimodales appliquées à TikTok (Balat *et al.*, 2024; Nguyen *et al.*, 2025), soit sur des architectures développées pour des plateformes de vidéos longues telles que YouTube (Ahmed *et al.*, 2024; Na *et al.*, 2024; Baharlouei *et al.*, 2024). Ces approches, bien qu'efficaces dans leur contexte respectif, ne prennent que partiellement en compte les spécificités structurelles des vidéos courtes fortement éditées et socialement contextualisées.

En particulier, sur TikTok, le caractère inapproprié d'un contenu ne se manifeste pas toujours par un indice visuel explicite. Dans ce mémoire, les contenus inappropriés désignent notamment les contenus violents, sexualisés, liés à l'automutilation ou au suicide. Il peut résulter d'une combinaison subtile entre le signal audiovisuel, les descriptions associées, les commentaires des utilisateurs et les dynamiques d'engagement. Cette interdépendance entre modalités rend les approches reposant sur une analyse isolée potentiellement insuffisantes, notamment dans le cas de contenus implicites, ambigus ou contextuellement sensibles.

Par ailleurs, si la littérature en apprentissage multimodal propose diverses stratégies de fusion, celles-ci demeurent souvent génériques et principalement orientées vers l'amélioration des performances globales. Elles analysent rarement la contribution différentielle des modalités ou la manière dont celles-ci interagissent dans un processus décisionnel contextuel. Or, dans le cadre de la modération de contenus destinés aux mineurs, comprendre le rôle relatif de chaque source d'information constitue un enjeu important tant du point de vue scientifique qu'éthique.

Ces limites soulignent la nécessité de développer une approche multimodale spécifiquement adaptée aux caractéristiques de TikTok, capable non seulement d'intégrer conjointement des signaux visuels, audio, textuels et comportementaux, mais également de modéliser explicitement leurs interactions et leur influence relative dans la détection de contenus inappropriés.

Face à ces limites, il devient nécessaire de proposer une approche mieux adaptée aux spécificités de TikTok.

0.3 OBJECTIF DE RECHERCHE

L'objectif principal de cette recherche est de proposer et d'évaluer une approche multimodale adaptée aux spécificités socio-contextuelles de TikTok pour la détection automatique de contenus inappropriés auxquels les enfants peuvent être exposés. Ce travail exploite conjointement plusieurs sources d'information disponibles dans les vidéos courtes visuelles, audio, textuelles et comportementales afin d'améliorer la robustesse et la fiabilité des systèmes de modération automatique. Plus précisément, ce mémoire poursuit les objectifs suivants :

— Concevoir une architecture multimodale capable d'intégrer efficacement les signaux visuels, audio, textuels (descriptions et commentaires) ainsi que les métadonnées d'engagement propres à TikTok ;

— Étudier des mécanismes de fusion multimodale permettant de moduler dynamiquement l'influence relative de chaque modalité tout en capturant leurs interactions ;

— Évaluer l'apport de chaque modalité dans la détection de contenus inappropriés à travers des expériences comparatives et des études d'ablation ;

— Analyser l'impact des choix architecturaux (fusion, mécanismes d'attention, Réseau LSTM bidirectionnel (Bidirectional Long Short-Term Memory) (BiLSTM)) sur les performances du modèle.

Afin de répondre à ces objectifs, ce travail apporte plusieurs contributions scientifiques et méthodologiques.

0.4 CONTRIBUTION DE LA RECHERCHE

Cette recherche apporte des contributions à la fois méthodologiques, expérimentales et applicatives à la problématique de la modération automatique sur TikTok. Les principales contributions sont les suivantes :

— L'enrichissement du dataset TikHarm par l'intégration des descriptions, des commentaires et des métadonnées d'engagement (likes, partages, nombre de commentaires), permettant de mieux refléter la nature multimodale et sociale des contenus TikTok ;

— La proposition de SoSAFE, une architecture multimodale spécifiquement conçue pour TikTok, intégrant cinq modalités complémentaires (vidéo, audio, description, commentaires et métadonnées) au sein d'un module de fusion combinant des mécanismes de gating par modalité, de cross-attention et un BiLSTM bidirectionnel ;

— La mise en évidence de l'apport des commentaires et des métadonnées, souvent négligés dans la littérature, pour la désambiguïsation de contenus implicites ou contextuellement sensibles et une évaluation expérimentale approfondie sur le dataset TikHarm enrichi, incluant une comparaison avec plusieurs approches de l'état de l'art dans un cadre expérimental homogène ;

CHAPITRE I

TRAVAUX CONNEXES

Ce chapitre présente un état de l'art des approches existantes pour la détection automatique de contenus inappropriés sur les plateformes de partage de vidéos. Il vise à analyser les principales contributions de la littérature, en mettant en évidence leurs apports et leurs limites, en particulier dans le contexte des contenus destinés aux enfants. Les travaux sont organisés en deux catégories principales : les approches unimodales, reposant sur une seule source d'information, et les approches multimodales, qui exploitent conjointement plusieurs modalités afin de mieux capturer la complexité des contenus.

1.1 APPROCHES UNIMODALES

Les approches unimodales constituent les premières tentatives visant à automatiser la détection de contenus inappropriés destinés aux enfants sur les plateformes de partage de vidéos. Elles reposent sur l'exploitation d'une seule modalité, le plus souvent la modalité visuelle, en raison de la nature intrinsèquement visuelle des contenus vidéo. Bien que ces méthodes aient permis d'établir des bases solides pour l'analyse automatique des contenus, elles présentent des limites importantes lorsqu'il s'agit de capturer des signaux implicites, contextuels ou multimodaux. Une part importante de la littérature s'est ainsi concentrée sur l'analyse visuelle à l'aide de réseaux neuronaux profonds. Ishikawa *et al.* (2019) proposent différentes architectures de réseaux neuronaux convolutifs pour identifier des dessins animés perturbants associés au phénomène Elsagate. Leur approche repose exclusivement sur les caractéristiques visuelles extraites des images, sans prise en compte de l'audio ou du texte. Bien que les résultats obtenus soient satisfaisants pour des contenus explicitement perturbants, cette approche demeure limitée face à des situations où le caractère nuisible est véhiculé de manière

implicite. Dans une logique similaire, Khan *et al.* (2024) introduisent une approche séquentielle basée sur un modèle Mémoire à long court terme (Long Short-Term Memory) (LSTM) pour la détection de scènes violentes dans les dessins animés avec un accuracy de 97%. Malgré la prise en compte de la dimension temporelle, la décision repose uniquement sur des caractéristiques visuelles, ce qui limite l'interprétation de contenus ambigus ou dépendants du contexte narratif. D'autres travaux, tels que ceux de Yousaf & Nawaz (2022), exploitent également des descripteurs visuels extraits de frames vidéo pour analyser des contenus YouTube destinés aux enfants. Bien que ces méthodes démontrent l'efficacité des informations visuelles, elles ignorent les signaux audio et textuels, réduisant leur robustesse face à des contenus subtils ou implicites. Singh *et al.* (2019), à travers le système KIDSGUARD, adoptent une approche strictement visuelle basée sur l'analyse des frames vidéo encodées par un réseau Visual Geometry Group 16 couches (VGG16), suivie d'un auto-encodeur LSTM. Leur approche atteint une précision de 80 % sur un jeu de données de 109 835 clips vidéo. Dans le contexte spécifique de TikTok, Balat *et al.* (2024) introduisent TikGuard, un modèle fondé sur des Transformers vidéo, évalué sur le dataset TikHarm. Bien que cette approche basée sur les Transformers (TimeSformer, VideoMAE, Video Vision Transformer (ViViT)) établisse une première référence sur TikTok, elle repose exclusivement sur la modalité visuelle, ce qui peut conduire à des erreurs lorsque l'information nuisible est portée par l'audio, le texte ou le contexte social.

Dans l'ensemble, ces approches montrent que l'analyse visuelle constitue un élément central pour la détection de contenus inappropriés, mais qu'elle reste insuffisante pour traiter des contenus implicites, ironiques ou fortement contextualisés.

1.2 APPROCHES MULTIMODALES

Face aux limites des approches unimodales, de nombreux travaux récents se sont orientés vers des stratégies multimodales visant à exploiter simultanément plusieurs sources d'information disponibles dans les vidéos en ligne. L'objectif est de capturer des indices complémentaires issus des modalités visuelle, audio, textuelle et contextuelle afin de mieux appréhender la complexité sémantique et sociale des contenus.

Parmi les premiers travaux, Tahir *et al.* (2019) proposent une architecture bimodale combinant les caractéristiques visuelles et acoustiques pour la détection de contenus inappropriés sur YouTube Kids, atteignant des performances supérieures à 90 % d'accuracy. Ces résultats montrent que l'intégration de l'audio constitue un apport significatif par rapport aux approches purement visuelles.

Des approches plus riches combinent également le texte et l'image. Chuttur & Nazurally (2022) exploitent conjointement des informations visuelles et textuelles issues des descriptions et sous-titres à travers LSTM et VGGNet pour analyser des dessins animés, atteignant environ 76 % d'accuracy sur des vidéos YouTube réelles. Bien que cette approche démontre l'intérêt du texte, la stratégie de fusion demeure relativement simple et ne modélise pas explicitement les interactions complexes entre modalités.

Plus récemment, plusieurs travaux proposent des architectures trimodales intégrant la vidéo, l'audio et le texte. Na *et al.* (2024) introduisent un réseau Enhanced Multimodal Transformer Network (E-MTN) amélioré pour la détection de scènes violentes, rapportant une moyenne de Précision moyenne (Average Precision) (AP) de 84 % sur le dataset XD-Violence. D'autres approches comme celle de Ahmed *et al.* (2024) exploitent des modèles multimodaux de grande capacité, notamment Pré-entraînement contrastif image–texte (Contrastive Language–Image Pre-training) (CLIP) et AudioCLIP enrichis par des prompts textuels, afin

d'aligner les représentations multimodales avec des descriptions textuelles des classes, obtenant des gains notables en robustesse et en capacité de généralisation avec 81.49% d'accuracy sur le dataset Multimodal Malicious Or Benign (MMOB) . Par ailleurs Baharlouei *et al.* (2024) introduisent Hierarchical Cross-modal Context-Aware Pooling (HICCAP), une architecture multimodale hiérarchique fondée sur des mécanismes de cross-attention entre le texte, la vidéo et l'audio. Le texte est encodé à l'aide de BERT, la video via un réseau I3D et l'audio par VGGish. La cross-attention permet de modéliser explicitement les interactions entre les modalités, en mettant en évidence les dépendances contextuelles entre le contenu visuel, les signaux acoustiques et les indices textuels. Cette approche se révèle particulièrement pertinente pour la détection des contenus ambigus, mêlant humour, violence ou sexualité implicite avec un F1-score de 74.96%. Des travaux plus récents se sont explicitement concentrés sur la plateforme TikTok. Nguyen *et al.* (2025) étendent TikGuard vers un cadre multimodal dédié, en combinant des caractéristiques visuelles avec des signaux textuels automatiquement extraits de la vidéo via à la Reconnaissance optique de caractères (Optical Character Recognition) (OCR) et la Reconnaissance automatique de la parole (Automatic Speech Recognition) (ASR). Les performances rapportées sur TikHarm, avec une accuracy et un F1-score macro proches de 89 %, confirment le rôle central du texte présent dans la vidéo pour la modération automatique sur TikTok. Toutefois, cette approche reste limitée aux signaux directement issus du contenu vidéo, sans prise en compte explicite du contexte social.

Par ailleurs, Singh *et al.* (2025) proposent une stratégie de fusion multimodale hiérarchique intégrant vision, audio et texte. Leurs résultats mettent en évidence l'intérêt de mécanismes de fusion avancés pour améliorer la robustesse et la capacité de généralisation des modèles, en particulier face à des contenus ambigus ou faiblement informatifs dans une seule modalité.

Enfin, certains travaux intègrent des signaux comportementaux et contextuels. Binh *et al.* (2022), à travers le modèle Samba, montrent que l'utilisation conjointe du texte avec BERT et des métadonnées d'engagement avec unité récurrente à portes (Gated Recurrent Unit) (GRU) permet d'atteindre un Score F1 (moyenne harmonique de la précision et du rappel) (F1)-score de 95 % sur un large corpus de vidéos YouTube. Ces résultats soulignent l'intérêt des signaux sociaux pour la contextualisation des contenus, bien que les informations audiovisuelles ne soient pas exploitées.

Ce chapitre a permis de mettre en évidence les principales approches proposées dans la littérature pour la détection de contenus inappropriés, ainsi que leurs limites. Les approches unimodales, bien qu'efficaces pour des contenus explicites, peinent à capturer des signaux implicites et contextuels. Les approches multimodales offrent des améliorations significatives en combinant plusieurs sources d'information, mais restent souvent génériques et peu adaptées aux spécificités de TikTok.

Ces observations mettent en évidence la nécessité de concevoir des architectures multimodales capables d'intégrer des modalités hétérogènes, tout en prenant en compte le contexte dynamique propre aux vidéos courtes. Dans cette continuité, le Chapitre 2 présente la problématique de recherche ainsi que les questions et hypothèses qui guident ce travail.

CHAPITRE II
PROBLÉMATIQUE ET CADRE DE RECHERCHE

Ce chapitre présente la problématique de recherche ainsi que le cadre conceptuel dans lequel s'inscrit ce travail. À partir d'une synthèse critique de l'état de l'art, il met en évidence les limites des approches existantes pour la détection de contenus inappropriés sur les plateformes de vidéos courtes, en particulier TikTok. Sur cette base, la problématique scientifique est formulée, suivie des questions de recherche et des hypothèses qui guident la conception de l'approche proposée.

2.1 SYNTHÈSE CRITIQUE DE L'ÉTAT DE L'ART

Les travaux présentés dans l'état de l'art témoignent de progrès significatifs dans la détection automatique des contenus inappropriés destinés aux enfants sur les plateformes de partage de vidéos. Les premières approches, essentiellement unimodales, ont démontré que l'analyse visuelle fondée sur des réseaux neuronaux profonds permettait d'identifier efficacement des contenus explicitement perturbants, notamment dans des vidéos animées ou destinées à un jeune public. Toutefois, ces méthodes montrent leurs limites lorsque le caractère nuisible d'un contenu n'est pas directement visible dans l'image, mais véhiculé par des indices audio, linguistiques ou contextuels.

Les approches multimodales ont permis de dépasser partiellement ces limites en combinant plusieurs sources d'information telles que la vidéo, l'audio et le texte. Plusieurs études ont montré que la fusion audiovisuelle ou trimodale améliore la robustesse de la détection, en particulier face à des contenus violents ou ambigus. Néanmoins, la majorité de ces travaux se concentrent sur des plateformes comme YouTube ou YouTube Kids, caractérisées par des

vidéos plus longues, moins éditées et moins dépendantes de dynamiques sociales immédiates que celles observées sur TikTok.

Par ailleurs, l'intégration explicite des métadonnées d'engagement, mentions « j'aime », partages, commentaires demeure marginale dans la littérature. Or, sur TikTok, ces signaux comportementaux participent pleinement à la construction du sens et peuvent refléter la réception collective d'un contenu comme sensible ou problématique. De plus, les architectures de fusion proposées sont souvent génériques, fréquemment basées sur de simples concaténations ou sur des mécanismes d'attention peu adaptés à des séquences multimodales très courtes, où chaque modalité est résumée par un unique embedding.

Un autre défi majeur concerne la généralisation aux plateformes de vidéos courtes. De nombreuses études reposent sur des jeux de données issus de YouTube ou collectés manuellement, sans validation sur des environnements tels que TikTok, marqués par des contenus courts, fortement édités et inscrits dans des dynamiques culturelles évolutives. Cette spécificité limite le transfert direct des approches existantes.

Enfin, malgré des performances quantitatives élevées, plusieurs travaux soulignent la difficulté des modèles face aux contenus implicites, ironiques ou fortement contextualisés. L'absence d'interprétabilité ainsi que le manque de partage de code ou de données entravent également la reproductibilité scientifique. Ces limites mettent en évidence la nécessité de concevoir des approches multimodales spécifiquement adaptées aux plateformes de vidéos courtes, capables d'exploiter conjointement les signaux audiovisuels, textuels et comportementaux tout en modélisant explicitement leurs interactions.

2.2 PROBLÉMATIQUE DE RECHERCHE

Au regard des limites identifiées dans l'état de l'art, la détection automatique de contenus inappropriés sur TikTok soulève un défi spécifique : les approches existantes échouent à modéliser explicitement les interactions entre modalités, ce qui limite leur capacité à détecter des contenus implicites ou fortement dépendants du contexte.

La littérature montre que les méthodes unimodales ou les fusions multimodales génériques ne permettent pas de capturer pleinement la complexité décisionnelle associée à des contenus implicites, ambigus ou fortement contextualisés. La difficulté ne réside pas uniquement dans l'extraction de caractéristiques pertinentes, mais dans la manière d'articuler dynamiquement des signaux complémentaires afin de refléter un processus de modération proche des conditions réelles d'usage. Cette limite est particulièrement critique dans le contexte de TikTok, où la nocivité émerge souvent de l'interaction entre modalités. La problématique scientifique de ce mémoire peut ainsi être formulée comme suit : comment concevoir une architecture multimodale capable de modéliser explicitement les interactions entre signaux audiovisuels, textuels et comportementaux afin d'améliorer la détection de contenus inappropriés destinés aux enfants sur TikTok ?

2.3 QUESTIONS DE RECHERCHE

Dans le prolongement de la problématique présentée, ce travail vise à mieux comprendre comment exploiter efficacement la multimodalité dans le contexte spécifique de TikTok, où les contenus sont à la fois courts, implicites et fortement influencés par des dynamiques sociales. L'objectif est d'analyser le rôle des différentes sources d'information, ainsi que les mécanismes permettant de les combiner de manière pertinente pour améliorer la détection de contenus

inappropriés destinés aux enfants. Dans ce cadre, les questions de recherche suivantes sont formulées :

— QR1 : L'intégration des commentaires et des métadonnées d'engagement améliore-t-elle de manière significative les performances de détection par rapport aux approches purement audiovisuelles ?

— QR2 : Les mécanismes de fusion multimodale adaptative présentent-ils des avantages par rapport aux stratégies de fusion statiques ou naïves ?

2.4 CADRE ET HYPOTHÈSES DE RECHERCHE

Ce travail s'inscrit dans le cadre de l'apprentissage profond multimodal appliqué à la modération automatique de contenus destinés aux enfants. Il repose sur l'hypothèse générale que la détection fiable de contenus inappropriés sur TikTok nécessite une modélisation explicite des interactions entre signaux audiovisuels et sociaux. Les hypothèses principales sont les suivantes :

— H1 : l'intégration conjointe des signaux audiovisuels, des signaux textuels (descriptions et commentaires) et des métadonnées d'engagement améliore les performances par rapport aux approches unimodales ou partiellement multimodales.

— H2 : les mécanismes de fusion adaptative sont plus performants que les stratégies de pondération fixes dans le cas de contenus courts à forte composante sociale et contextuelle.

Ce chapitre a permis d'identifier les limites des approches existantes et de formuler une problématique de recherche adaptée aux spécificités de TikTok. Il met en évidence la nécessité de développer des méthodes capables de mieux modéliser les interactions entre des modalités hétérogènes ainsi que les informations contextuelles associées.

Les questions de recherche et les hypothèses définissent un cadre structurant pour l'ensemble du travail et orientent les choix méthodologiques adoptés. Dans cette continuité, le Chapitre 3 présente la méthodologie mise en place pour répondre à ces enjeux.

CHAPITRE III
MÉTHODOLOGIE

Ce chapitre décrit la méthodologie adoptée pour la détection de contenus inappropriés sur TikTok. Il présente dans un premier temps le dataset TikHarm ainsi que son enrichissement multimodal, incluant l'intégration des descriptions, des commentaires et des métadonnées d'engagement. Les différentes étapes de prétraitement appliquées à chaque modalité sont ensuite détaillées, afin de garantir la cohérence et la qualité des données exploitées. Enfin, ce chapitre introduit les méthodes d'extraction de caractéristiques ainsi que l'architecture multimodale proposée, conçue pour modéliser efficacement les interactions entre les différentes sources d'information.

3.1 PRÉPARATION DE L'ENSEMBLE DE DONNÉES

Cette section présente le jeu de données TikHarm utilisé dans ce travail. Nous décrivons d'abord le dataset tel qu'il a été introduit dans la littérature, puis nous détaillons les classes et le processus d'annotation. Nous expliquons ensuite la version effectivement exploitée dans notre étude ainsi que la stratégie de partition adoptée. Enfin, nous discutons des principales limites du dataset, tant intrinsèques qu'opérationnelles.

3.1.1 DESCRIPTION GÉNÉRALE DU DATASET

TikHarm est un jeu de données annoté conçu pour la détection automatique de contenus inappropriés destinés aux enfants sur la plateforme TikTok. Il a été initialement introduit par Balat *et al.* (2024) dans le cadre de leur étude TikGuard, afin de répondre à l'absence de jeux de données publics spécifiquement dédiés à la modération de contenus sur TikTok.

Contrairement à des plateformes plus anciennes, TikTok se distingue par des vidéos de courte durée, fortement éditées, le plus souvent accompagnées de musique, d'effets visuels et d'interactions sociales immédiates. Ces caractéristiques rendent la détection automatique de contenus inappropriés particulièrement complexe et justifient l'utilisation d'un dataset spécifiquement conçu pour ce contexte. À notre connaissance et selon les travaux recensés dans la littérature récente, TikHarm constitue à ce jour le seul jeu de données public explicitement dédié à la modération de contenus TikTok, avec une annotation centrée sur le risque pour les mineurs. Le dataset comprend initialement 3 948 vidéos TikTok, collectées à partir d'Uniform Resource Locator (URL) publiques accessibles aux enfants et annotées manuellement selon leur caractère approprié ou non pour un public mineur. Chaque vidéo est associée à une étiquette unique correspondant à une catégorie de contenu. TikHarm vise à couvrir une diversité de situations, allant de contenus explicitement nuisibles à des vidéos apparemment anodines mais susceptibles de présenter des risques implicites pour les enfants. Ainsi, TikHarm constitue l'un des premiers jeux de données structurés spécifiquement pour l'analyse de contenus inappropriés sur TikTok et sert aujourd'hui de référence pour l'évaluation de modèles de détection automatique sur cette plateforme.

3.1.2 CLASSES ET PROCESSUS D'ANNOTATION

Le dataset TikHarm est structuré autour de quatre classes distinctes, correspondant à différents niveaux de risque pour les jeunes utilisateurs. La classe Harmful Content regroupe des vidéos présentant de la violence ou des comportements dangereux susceptibles d'être imités par les enfants. La classe Adult Content concerne les contenus à caractère sexuel ou tout autre matériel jugé inapproprié pour un public mineur. La classe Safe inclut des vidéos considérées comme non dangereuses, telles que des contenus éducatifs ou des dessins animés. Enfin, la classe Suicide regroupe les vidéos abordant explicitement ou implicitement des

comportements ou des idées suicidaires. La distribution de ces différentes classes dans le dataset est présentée dans le Tableau 3.1.

TABLEAU 3.1 : Distribution des classes.

		Duration (s)		
Classe	**Exemples**	**Min**	**Max**	**Moyenne**
Adult Content	768	1.79	442.94	34.84
Harmful Content	921	4.80	600.00	36.41
Safe	874	5.15	568.79	65.11
Suicide	848	4.03	429.50	18.02

L'annotation du dataset a été réalisée manuellement par des annotateurs formés, suivant des consignes précises afin d'assurer une classification cohérente et fiable des vidéos dans chacune des catégories. Ce processus vise à garantir la qualité des étiquettes utilisées pour l'entraînement et l'évaluation des modèles de classification automatique (Balat *et al.*, 2024).

3.1.3 LIMITES DU DATASET

Malgré son intérêt et sa pertinence, le dataset TikHarm présente plusieurs limites qu'il convient de souligner. Premièrement, la taille du jeu de données demeure relativement modeste au regard de l'échelle réelle de TikTok, où des millions de vidéos sont publiées quotidiennement. Cette contrainte peut limiter la capacité des modèles à généraliser à des contenus très variés ou rares. Deuxièmement, la disponibilité des vidéos constitue un défi majeur. Les vidéos TikTok peuvent être supprimées ou rendues privées, ce qui affecte la reproductibilité des expériences et la stabilité du dataset dans le temps. Troisièmement, l'annotation humaine, bien que nécessaire, introduit une part de subjectivité. Les frontières entre certaines classes, notamment entre Harmful Content et Suicide, peuvent être floues lorsque les contenus sont implicites, ironiques ou contextuels. Par ailleurs, la majorité des vidéos du dataset sont en vietnamien, ce qui peut limiter la généralisation des modèles à d'autres contextes linguistiques

et culturels. Enfin, dans sa version originale, TikHarm ne fournit pas l'ensemble des modalités exploitables (commentaires, métadonnées d'engagement), ce qui limite son utilisation pour des approches multimodales avancées. Ces limites motivent l'enrichissement du dataset présenté dans la section suivante, qui constitue l'une des contributions principales de ce travail.

3.1.4 ENRICHISSEMENT DU DATASET TIKHARM

Le jeu de données TikHarm, dans sa version originale, repose principalement sur le contenu vidéo et son annotation manuelle. Bien que cette base soit pertinente, elle ne permet pas de capturer l'ensemble des signaux disponibles sur TikTok, une plateforme intrinsèquement multimodale et sociale. En particulier, les informations textuelles associées aux vidéos (descriptions et commentaires), ainsi que les métadonnées d'engagement, constituent des sources d'information susceptibles d'apporter un contexte essentiel à l'interprétation des contenus. Dans ce travail, nous avons procédé à un enrichissement du dataset TikHarm afin d'y intégrer ces modalités supplémentaires. Cet enrichissement vise à renforcer la dimension multimodale du dataset, à mieux refléter les usages réels de TikTok et à permettre l'apprentissage de modèles capables d'exploiter conjointement des signaux visuels, audio, textuels et comportementaux. La collecte a été réalisée au mois de mai 2025.

DESCRIPTIONS DES VIDÉOS

La description associée à une vidéo TikTok constitue une première source d'information textuelle fournie par le créateur du contenu. Elle peut inclure du texte libre, des hashtags, des emojis, ainsi que des mentions d'autres utilisateurs ou de tendances spécifiques. Ces éléments jouent un rôle important dans la contextualisation de la vidéo et dans la transmission de l'intention communicative de son auteur.

Pour chaque vidéo du dataset exploité, nous avons reconstruit l'URL correspondante à partir de l'identifiant de l'utilisateur. À l'aide d'un pipeline automatisé basé sur Selenium WebDriver, nous avons ensuite extrait la description textuelle complète lorsqu'elle était disponible. Celle-ci peut contenir des indices explicites, tels que des mots-clés liés à la violence, à la sexualité ou à des thématiques sensibles, mais également des indices implicites, par exemple sous forme d'ironie, de sous-entendus ou de références culturelles.

Les descriptions TikTok présentent toutefois plusieurs défis. Elles sont souvent courtes et peuvent contenir du bruit, ce qui limite la richesse de l'information textuelle exploitable. De plus, elles sont fréquemment multilingues ou comportent des fautes d'orthographe, ce qui complique leur traitement automatique. Enfin, elles intègrent régulièrement des emojis et des hashtags, qui véhiculent une forte charge sémantique mais restent difficiles à interpréter de manière fiable par les modèles.

Malgré ces limites, l'intégration des descriptions permet d'enrichir significativement la représentation des vidéos en apportant un point de vue textuel complémentaire aux signaux visuels et audio.

COMMENTAIRES DES UTILISATEURS

Les commentaires constituent une modalité particulièrement intéressante, car ils reflètent les réactions de la communauté face à une vidéo. Contrairement à la description, les commentaires sont produits par les spectateurs et peuvent révéler la manière dont le contenu est perçu, interprété ou détourné.

En utilisant le même pipeline de collecte que celui décrit pour l'extraction des descriptions, l'ensemble des commentaires visibles associés à chaque vidéo a été récupéré. Ces

commentaires ont ensuite été triés selon le nombre de mentions « j'aime », et seuls les K=10 commentaires les plus appréciés ont été conservés. Ce choix repose sur plusieurs considérations. D'une part, les commentaires les plus appréciés sont généralement les plus représentatifs de la polarisation des réactions des utilisateurs. D'autre part, ils tendent à contenir des signaux émotionnels forts, tels que l'approbation, l'indignation, l'humour ou l'inquiétude. Enfin, cette sélection permet de limiter le bruit en écartant les commentaires non pertinents ou hors sujet. Ce choix permet également de maintenir une taille d'entrée textuelle compatible avec les contraintes computationnelles du modèle. Les commentaires peuvent jouer un rôle crucial dans la détection de contenus inappropriés. En effet, une vidéo visuellement neutre peut susciter des réactions alarmantes ou critiques, révélant un caractère problématique implicite. À l'inverse, certains contenus explicitement choquants peuvent être recontextualisés à travers des commentaires de prévention ou de dénonciation. L'intégration des commentaires permet ainsi de prendre en compte une dimension sociale et interprétative qui reste absente des approches reposant uniquement sur les signaux visuels ou audiovisuels.

MÉTADONNÉES D'ENGAGEMENT

En complément des contenus multimodaux, TikTok fournit un ensemble de métadonnées d'engagement associées à chaque vidéo. Dans ce travail, nous avons extrait plusieurs indicateurs à l'aide du même pipeline que celui utilisé pour les modalités textuelles, complété par l'Interface de Programmation d'Applications (API) TikTok Scraper de la plateforme Apify (Apify, 2025). Ces indicateurs incluent notamment le nombre de mentions « j'aime » (likes), le nombre de partages (shares), ainsi que le nombre total de commentaires. Ces indicateurs fournissent une information quantitative sur la popularité, la viralité et la résonance sociale d'une vidéo. Des travaux ont montré que certains types de contenus sensibles ou controversés tendent à générer des niveaux d'engagement spécifiques, caractérisés par une forte polarisation

ou une diffusion rapide (Bonifazi *et al.*, 2024). Les métadonnées d'engagement présentent toutefois un double aspect. D'une part, elles peuvent constituer des signaux discriminants utiles pour la classification. D'autre part, elles sont sensibles à des facteurs externes tels que les tendances, les effets de mode ou les mécanismes de recommandation, et doivent donc être interprétées avec prudence. Dans notre approche, ces métadonnées sont considérées comme des signaux comportementaux complémentaires, venant enrichir les représentations visuelles, audio et textuelles, sans s'y substituer.

L'enrichissement du dataset TikHarm permet de transformer un jeu de données initialement centré sur la vidéo en un dataset véritablement multimodal, intégrant :

— le contenu visuel et audio des vidéos,

— le texte produit par le créateur (description),

— le texte produit par la communauté (commentaires),

— des indicateurs d'engagement social (métadonnées).

Cet enrichissement constitue une étape clé de notre méthodologie et conditionne la conception de l'architecture multimodale proposée, présentée dans les sections suivantes. Il permet également de mieux capturer la complexité et la dimension contextuelle des contenus TikTok, en particulier lorsque les signaux nuisibles sont implicites ou distribués entre plusieurs modalités. Il conditionne directement la conception et l'évaluation du modèle proposé

3.1.5 TAILLE DU DATASET ET PARTITION DES DONNÉES

Bien que TikHarm fournisse une version initiale du dataset, celle-ci n'a pas été utilisée telle quelle dans notre travail. En pratique, plusieurs vidéos référencées dans la version originale n'étaient plus accessibles au moment de nos expérimentations, et notre approche repose sur un enrichissement multimodal du dataset, incluant l'ajout des descriptions, des

commentaires et des métadonnées d'engagement. Cet enrichissement nécessite que l'ensemble des modalités soit disponible de manière cohérente pour chaque vidéo, ce qui a conduit à l'exclusion de certains échantillons incomplets. La version finale exploitée dans notre étude comprend ainsi 3 411 vidéos, réparties entre les quatre classes du dataset. Sur cette base, nous avons construit une partition spécifique, distincte de celle éventuellement utilisée dans les travaux initiaux. Le dataset a été divisé en deux sous-ensembles disjoints, avec 90 % des données dédiées à l'entraînement et 10 % réservées à l'évaluation. Il est ainsi composé d'un ensemble d'apprentissage (train) et d'un ensemble de test (test). La partition a été réalisée de manière stratifiée afin de préserver la distribution des classes dans chaque sous-ensemble, ce qui permet de garantir une évaluation plus fiable des performances tout en limitant les biais liés au déséquilibre de classes. Les proportions retenues s'inscrivent dans les pratiques courantes en apprentissage automatique supervisé. La répartition des données entre les différents sous-ensembles est présentée dans le Tableau 3.2.

TABLEAU 3.2 : Distribution des sous-ensembles.

		Durée (s)		
Partition	**Exemples**	**Min**	**Max**	**Moyenne**
Train	3069	5.94	892.11	75.12
Test	342	5.15	540.25	41.53

L'ensemble d'apprentissage est utilisé dans le cadre d'une validation croisée stratifiée en 5 plis, décrite en détail dans la Section 4.2 intitulée Protocole expérimental. Toutes les expériences menées dans ce travail, ainsi que les baselines reproduites ou adaptées, sont évaluées sur cette même partition afin d'assurer une comparaison équitable.

3.2 PRÉTRAITEMENT DES DONNÉES

Les données issues de TikTok présentent une forte hétérogénéité, tant au niveau des formats que de la qualité des informations disponibles. Avant toute phase d'extraction de

caractéristiques et d'apprentissage, un prétraitement rigoureux est indispensable afin d'assurer la cohérence des entrées, de réduire le bruit et de rendre les différentes modalités exploitables par les modèles d'apprentissage automatique. Dans ce travail, un prétraitement spécifique est appliqué à chaque modalité vidéo, audio, texte et métadonnées en tenant compte de leurs caractéristiques propres ainsi que des contraintes imposées par les modèles utilisés par la suite.

3.2.1 PRÉTRAITEMENT DES DONNÉES VIDÉO

La vidéo constitue la modalité principale du jeu de données TikHarm. Toutefois, les vidéos issues de TikTok présentent une forte hétérogénéité en termes de durée, de résolution, de format d'encodage et de qualité visuelle. Afin de garantir une représentation homogène et exploitable par les modèles d'apprentissage profond, un pipeline de prétraitement vidéo standardisé a été mis en place.

Chaque vidéo est d'abord décodée à l'aide d'outils robustes et reproductibles. Selon la disponibilité des bibliothèques, le décodage est assuré prioritairement par Decord, puis par VideoReader de torchvision, ou à défaut par la fonction read-video. Cette stratégie garantit une extraction fiable des frames, indépendamment du format ou du codec de la vidéo.

Étant donné la durée généralement courte des vidéos TikTok, un échantillonnage temporel uniforme est appliqué afin d'extraire un nombre fixe de frames par vidéo. Dans notre configuration, 16 images sont sélectionnées de manière équidistante sur l'ensemble de la séquence. Lorsque la vidéo contient moins de 16 frames, certaines images sont répliquées afin de maintenir une taille d'entrée constante. Cette approche permet de couvrir l'évolution temporelle du contenu tout en maîtrisant le coût computationnel.

Les frames extraites sont ensuite redimensionnées de manière à conserver le ratio d'aspect, en ajustant le plus petit côté à 256 pixels, puis recadrées au centre pour obtenir une

résolution fixe de 224 × 224 pixels, compatible avec les architectures convolutionnelles et spatio-temporelles pré-entraînées. Les valeurs des pixels sont ensuite normalisées dans l'intervalle [0,1], puis standardisées à l'aide de moyennes et écarts-types prédéfinis correspondant aux statistiques du dataset ImageNet, assurant ainsi la compatibilité avec les poids pré-entraînés du modèle Réseau convolutionnel 3D inflaté (Inflated 3D Convolutional Network) (I3D).

Ce prétraitement vise à réduire la variabilité visuelle non pertinente liée à la résolution ou au cadrage, tout en conservant les informations spatiales et temporelles essentielles à la détection de contenus inappropriés.

3.2.2 PRÉTRAITEMENT DES DONNÉES AUDIO :

La piste audio des vidéos TikTok constitue une source d'information essentielle, notamment pour la détection de contenus explicites, de paroles sensibles ou d'indices émotionnels. Le prétraitement audio débute par l'extraction de la bande sonore associée à chaque vidéo.

L'audio est extrait à l'aide de l' Outil multimédia de traitement audio et vidéo (ffmpeg), puis converti en un signal mono échantillonné à 16 kHz, conformément aux paramètres requis par les modèles audio utilisés par la suite. Cette étape permet d'uniformiser les formats audio, indépendamment des caractéristiques initiales des vidéos (codec, fréquence d'échantillonnage ou nombre de canaux).

À partir du signal audio brut, des spectrogrammes log-Mel sont calculés avec **64 bandes** ($n_{mels} = 64$), conformément à la configuration du modèle VGGish utilisé pour l'extraction des représentations audio. Ces représentations temps-fréquence sont largement employées en traitement du signal audio, car elles capturent efficacement les propriétés perceptuelles du son. Les paramètres retenus incluent un nombre fixe de bandes de Mel, une fenêtre temporelle

courte et un pas constant, assurant une représentation cohérente et comparable entre les différentes vidéos.

Lorsque la durée du signal audio dépasse la fenêtre d'analyse maximale du modèle, celui-ci est découpé en segments de durée fixe correspondant à la taille d'entrée attendue par VGGish. Chaque segment est traité indépendamment, puis les représentations obtenues sont agrégées par moyenne afin de produire une représentation audio globale par vidéo. Cette stratégie permet de gérer des durées audio variables tout en conservant une taille d'entrée fixe pour l'extraction de caractéristiques.

3.2.3 PRÉTRAITEMENT DES DONNÉES TEXTUELLES

Les données textuelles exploitées dans ce travail proviennent de deux sources distinctes : les descriptions des vidéos fournies par les créateurs de contenu et les commentaires des utilisateurs. Ces données présentent un niveau de bruit élevé, caractéristique des réseaux sociaux, et nécessitent un prétraitement spécifique avant leur exploitation par des modèles de langage. Dans un premier temps, les textes sont normalisés par canonisation Unicode (NFKC) afin d'uniformiser les caractères et de réduire les variations liées à l'encodage. Les URLs, les espaces superflus ainsi que certains symboles non informatifs sont supprimés, tout en conservant les éléments porteurs de sens sémantique. Une attention particulière est portée au traitement des emojis, hashtags et mentions, omniprésents sur TikTok et fortement porteurs d'information contextuelle. Les emojis sont conservés et enrichis par leur transcription textuelle (demojization), permettant de préserver à la fois l'information visuelle et sa signification linguistique. Des catégories sémantiques générales (par exemple émotion, violence ou danger) sont également associées à certains emojis afin de renforcer leur interprétabilité par les modèles de langage. Les hashtags et mentions sont explicitement marqués et séparés du texte brut, afin de faciliter leur prise en compte par les modèles Transformer. Les descriptions et les

commentaires sont ensuite tronqués ou complétés par padding afin de respecter une longueur maximale compatible avec les architectures de type Transformer utilisées dans ce travail. Ce prétraitement vise à préserver la richesse sémantique des textes, tout en réduisant l'impact du bruit, de la variabilité linguistique et des artefacts propres aux réseaux sociaux, afin de fournir des entrées textuelles cohérentes et exploitables pour l'apprentissage multimodal.

3.2.4 PRÉTRAITEMENT DES MÉTADONNÉES D'ENGAGEMENT

Les métadonnées d'engagement, telles que le nombre de mentions « j'aime », de partages et de commentaires, constituent des variables numériques susceptibles de présenter une forte dispersion ainsi que des distributions fortement asymétriques, caractéristiques des dynamiques de popularité sur les réseaux sociaux. Afin de les rendre exploitables par des modèles d'apprentissage automatique, un prétraitement spécifique est appliqué.

Dans un premier temps, les valeurs manquantes ou non valides sont identifiées et traitées de manière cohérente. Les valeurs textuelles éventuellement présentes (par exemple sous la forme k ou m) sont converties en valeurs numériques continues. Les valeurs manquantes sont ensuite imputées à l'aide de la médiane, ce choix permettant de limiter l'influence des valeurs extrêmes.

Compte tenu de la distribution fortement déséquilibrée de ces variables, une transformation logarithmique de type log1p est appliquée afin de réduire l'effet des valeurs extrêmes et de stabiliser les distributions. Les métadonnées sont ensuite normalisées à l'aide d'un redimensionnement robuste (RobustScaler), qui repose sur la médiane et l'intervalle interquartile. Cette stratégie permet d'obtenir des variables comparables entre elles tout en étant moins sensibles aux outliers qu'une standardisation classique de type z-score.

Ce prétraitement permet d'intégrer les métadonnées comme des caractéristiques comportementales stables et informatives, facilitant leur fusion ultérieure avec les autres modalités visuelles, audio et textuelles au sein de l'architecture multimodale proposée.

L'ensemble de ces étapes garantit que les différentes modalités sont représentées dans des formats homogènes, compatibles avec les architectures pré-entraînées utilisées ultérieurement, tout en minimisant l'introduction de biais liés aux variations de format ou de qualité des données brutes. La section suivante détaille les modèles et méthodes utilisés pour extraire des représentations discriminantes à partir de ces données prétraitées.

3.3 EXTRACTION DES CARACTÉRISTIQUES

Après les étapes d'enrichissement et de prétraitement, chaque modalité est transformée en une représentation vectorielle compacte et discriminante, appelée embedding. L'objectif de cette phase est d'extraire, pour chaque vidéo, des caractéristiques pertinentes permettant de capturer à la fois le contenu explicite et les signaux contextuels implicites associés. Dans ce travail, nous adoptons une approche modulaire : chaque modalité est traitée par un modèle spécialisé, généralement pré-entraîné sur de larges corpus, puis partiellement adapté à la tâche de détection de contenus inappropriés. Cette stratégie permet de bénéficier de connaissances préalables tout en limitant le coût d'entraînement et le risque de sur-apprentissage.

3.3.1 EXTRACTION DES CARACTÉRISTIQUES VISUELLES

La modalité visuelle est représentée par les frames extraites lors de la phase de prétraitement. Les séquences d'images sont encodées à l'aide d'un réseau de type I3D (*Inflated 3D ConvNet*) (Carreira & Zisserman, 2017), instancié ici sous la forme `i3d_r50` proposée par PyTorchVideo, dont le backbone repose sur une architecture ResNet-50 3D (He *et al.*, 2016).

Ce type de modèle est particulièrement adapté à l'analyse vidéo, car il permet de capturer conjointement les informations spatiales (apparence des scènes) et temporelles (évolution du contenu au fil du temps). Le réseau est initialisé avec des poids pré-entraînés sur des jeux de données vidéo à grande échelle et affiné partiellement. Ce qui permet de bénéficier de représentations visuelles riches et généralisables. Afin de limiter le sur-apprentissage et de réduire le coût computationnel, la majorité du backbone est gelée durant l'entraînement. Seul le dernier stage du réseau, ainsi que la tête de classification, sont rendus entraînables. Cette stratégie de fine-tuning partiel permet d'adapter efficacement le modèle aux spécificités des vidéos TikTok, tout en conservant les connaissances visuelles générales acquises lors du pré-entraînement. Les activations produites par le réseau sont ensuite agrégées à l'aide d'un pooling spatio-temporel global de type adaptive average pooling, aboutissant à un vecteur de caractéristiques unique par vidéo. Ce vecteur encode des informations visuelles de haut niveau, telles que la présence de scènes violentes, de gestes inappropriés ou de signaux visuels ambigus, susceptibles d'indiquer un contenu nuisible ou sensible pour un public mineur.

3.3.2 EXTRACTION DES CARACTÉRISTIQUES ACOUSTIQUES

Les caractéristiques audio sont extraites à l'aide du modèle VGGish de (Hershey *et al.*, 2017), à partir des spectrogrammes log-Mel calculés lors de la phase de prétraitement. VGGish est une architecture convolutionnelle conçue pour l'analyse et la classification audio à grande échelle. Initialisé avec des poids pré-entraînés sur de vastes corpus audio, il est capable de capturer des motifs acoustiques généraux tels que la parole, la musique, les cris ou des sons brusques. Dans notre approche, le modèle VGGish est partiellement affiné afin de l'adapter aux spécificités des vidéos TikTok. L'ensemble des couches convolutionnelles du backbone est gelé, tandis que la dernière couche linéaire est rendue entraînable et couplée à une tête de classification lors de la phase de fine-tuning. Cette stratégie permet de préserver

les représentations acoustiques générales tout en adaptant le modèle aux signaux audio caractéristiques des contenus TikTok, tels que les voix modifiées, les musiques virales ou les effets sonores courts. Ces caractéristiques permettent de capturer des indices liés à l'émotion, à l'intensité sonore ou à des signaux auditifs implicites, souvent associés à des contenus potentiellement nuisibles. Lorsque la piste audio d'une vidéo est découpée en plusieurs segments temporels, un embedding est extrait pour chaque segment. Ces embeddings sont ensuite agrégés par moyenne, afin d'obtenir une représentation audio globale par vidéo. Le vecteur final, de dimension 128, encode des informations liées au ton, à l'intensité sonore et à la présence de signaux auditifs potentiellement sensibles ou émotionnellement marqués.

3.3.3 EXTRACTION DES CARACTÉRISTIQUES TEXTUELLES

Les données textuelles exploitées dans ce travail comprennent les descriptions des vidéos ainsi que les commentaires des utilisateurs. Ces deux sources sont traitées à l'aide d'un modèle de langage de type Transformer, à savoir BERT multilingue (Multilingual Bidirectional Encoder Representations from Transformers) (mBERT) de Devlin *et al.* (2019), afin de prendre en compte la diversité linguistique présente sur la plateforme TikTok. Les descriptions sont encodées individuellement par le modèle mBERT. La représentation textuelle finale est obtenue à partir des représentations contextuelles produites par le modèle, en appliquant un pooling moyen masqué (mean pooling) sur les sorties du dernier encodeur. Cette opération permet de produire un vecteur dense de dimension 768, représentant de manière globale le contenu sémantique de la description. Les commentaires sont traités séparément : chaque commentaire retenu est encodé individuellement à l'aide du même backbone mBERT. Étant donné le caractère multiple, hétérogène et bruité des commentaires, leurs représentations ne sont pas simplement moyennées. À la place, un mécanisme d'attention appris est introduit afin d'agréger dynamiquement les embeddings des commentaires. Ce mécanisme permet de

pondérer l'importance relative de chaque commentaire, en mettant en avant ceux qui apportent le plus d'information discriminante pour la détection de contenus inappropriés. Lors de la phase de fine-tuning, la majorité des couches du modèle de langage est gelée afin de préserver les connaissances linguistiques générales acquises lors du pré-entraînement. Seules les 4 dernières couches de l'encodeur mBERT, ainsi que le module d'attention sur les commentaires, la projection et la tête de classification, sont rendues entraînables. Cette stratégie permet de limiter le sur-apprentissage, tout en adaptant efficacement les représentations textuelles aux spécificités du langage utilisé sur TikTok, caractérisé par des formulations courtes, implicites et fortement contextuelles.

3.3.4 EXTRACTION DES CARACTÉRISTIQUES DES MÉTADONNÉES

Les métadonnées d'engagement prétraitées sont transformées en un vecteur de caractéristiques tabulaires, décrivant les dynamiques de popularité et de diffusion associées à chaque vidéo. Afin de modéliser des relations non linéaires et des interactions complexes entre ces variables, les métadonnées sont projetées à l'aide d'un réseau de neurones multicouches (Perceptron multicouches (Multilayer Perceptron) (MLP)) de faible profondeur. Avant la phase de projection, les variables numériques normalisées sont enrichies par l'application de transformations supplémentaires, incluant notamment des termes polynomiaux de faible degré, des indicateurs de valeurs nulles, des interactions croisées entre variables, ainsi que des opérations de binning. Ces transformations visent à mieux capturer les comportements atypiques ou extrêmes, fréquemment observés dans les distributions d'engagement sur les réseaux sociaux. Le MLP encode ensuite ces caractéristiques enrichies dans un embedding dense de dimension 32, représentant de manière compacte les dynamiques comportementales associées à chaque vidéo. Cette représentation capture des informations telles que la viralité, la polarisation des réactions ou l'intensité de l'engagement suscité par le contenu. Ces carac-

téristiques comportementales complètent les informations visuelles, audio et textuelles, en fournissant un signal contextuel indirect sur la réception et l'interprétation des contenus par la communauté des utilisateurs.

À l'issue de cette étape, chaque vidéo est représentée par un ensemble d'embeddings multimodaux correspondant aux différentes sources d'information :

— un embedding visuel,

— un embedding audio,

— un embedding textuel issu de la description,

— un embedding textuel agrégé issu des commentaires,

— un embedding issu des métadonnées d'engagement.

Ces représentations constituent les entrées du module de fusion multimodale, décrit dans la section suivante. Cette extraction modulaire favorise la séparation fonctionnelle des encodeurs et facilite l'analyse ablationnelle de la contribution relative de chaque modalité.

3.4 ARCHITECTURE MULTIMODALE PROPOSÉE

Après l'extraction des caractéristiques propres à chaque modalité, l'enjeu principal consiste à combiner efficacement ces informations hétérogènes afin de produire une décision de classification robuste. Les contenus TikTok étant intrinsèquement multimodaux, les approches basées sur une simple concaténation des embeddings restent insuffisantes, car elles ne permettent pas de modéliser explicitement les interactions complexes entre signaux visuels, audio, textuels et comportementaux.

Dans ce travail, nous proposons une architecture multimodale unifiée conçue pour projeter les représentations issues de différentes modalités dans un espace latent commun, tout en adaptant dynamiquement l'importance de chaque modalité en fonction du contexte de

la vidéo analysée. Contrairement aux approches classiques, notre méthode vise à modéliser explicitement les interactions inter-modales avant la phase de prédiction, afin de mieux capturer la nature contextuelle et implicite des contenus. Une vue d'ensemble de l'architecture proposée est présentée dans la Figure 3.1.

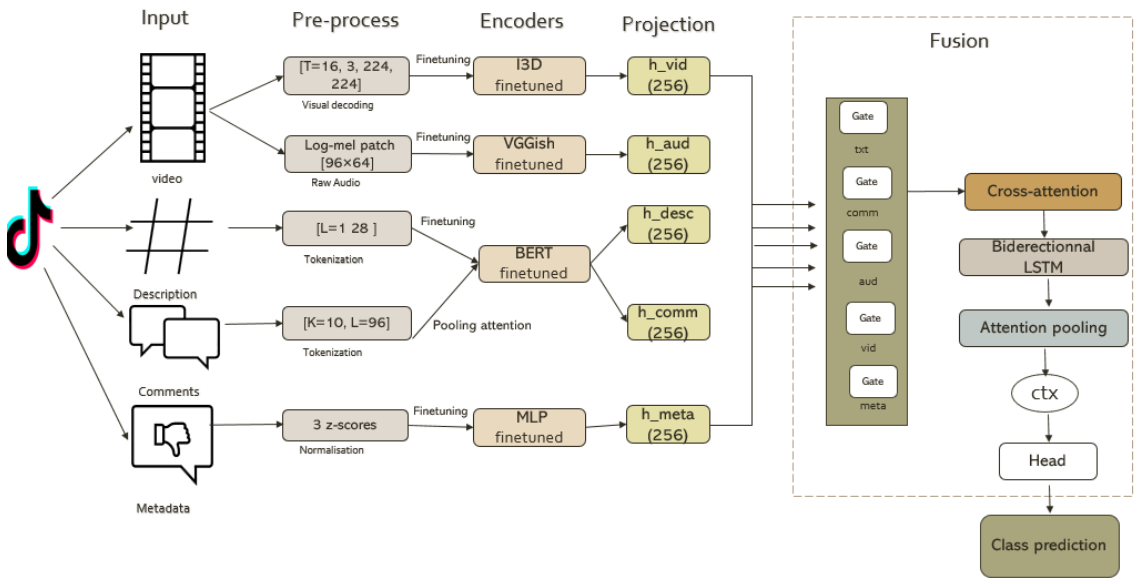

FIGURE 3.1 : Architecture proposée.

3.4.1 VUE D'ENSEMBLE DE L'ARCHITECTURE

Le processus de fusion repose sur plusieurs mécanismes complémentaires. Les commentaires sont d'abord agrégés à l'aide d'un mécanisme d'attention, permettant d'extraire une représentation unique et informative à partir des K commentaires sélectionnés. Un mécanisme de pondération dynamique par modalité est ensuite appliqué afin d'ajuster l'importance relative de chaque source d'information en fonction du contenu analysé.

Puis, une étape d'attention croisée inter-modale est introduite sur un graphe d'interactions parcimonieux, ce qui permet d'enrichir chaque modalité à partir des autres et de mieux capturer leurs dépendances. Enfin, une modélisation séquentielle basée sur un BiLSTM bidirectionnel, combinée à un mécanisme d'agrégation attentionnelle, permet d'obtenir une représentation globale de la vidéo. Cette représentation est ensuite exploitée par une tête de classification de type MLP pour produire les logits associés aux différentes classes du dataset TikHarm.

3.4.2 PROJECTION DANS UN ESPACE LATENT COMMUN

Les embeddings extraits à partir des différentes modalités présentent des dimensions hétérogènes, par exemple 2048 pour la vidéo, 128 pour l'audio et 768 pour les données textuelles. Afin de faciliter leur fusion, chaque modalité est projetée dans un espace latent commun de dimension d_{fuse}, dont la valeur est déterminée par optimisation des hyperparamètres. Concrètement, cette projection repose sur des couches linéaires suivies de fonctions de non-linéarité et de mécanismes de normalisation, ce qui permet d'obtenir des représentations homogènes et comparables entre les différentes modalités. Cette étape permet d'uniformiser les dimensions des différentes représentations, de limiter la domination des modalités de grande dimension, et d'introduire une première transformation adaptée à la tâche de classification. Elle constitue ainsi une étape clé pour stabiliser l'apprentissage et permettre une intégration cohérente des différentes sources d'information dans le processus de fusion multimodale.

3.4.3 AGRÉGATION ATTENTIVE DES COMMENTAIRES

Contrairement aux autres modalités représentées par un seul vecteur, les commentaires sont disponibles sous la forme d'un ensemble de K embeddings, chacun correspondant à

un commentaire. Afin d'obtenir une représentation unique, un mécanisme d'agrégation par attention est appliqué. Celui-ci consiste à attribuer un score à chaque commentaire, en tenant compte d'un masque indiquant leur présence effective, puis à calculer une combinaison linéaire pondérée de ces représentations. Cette approche permet de mettre en avant les commentaires les plus informatifs pour la tâche de classification.

Soit $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$, où $\mathbf{h}_k \in \mathbb{R}^d$ désigne la représentation du k-ième commentaire, et soit $\mathbf{m} \in \{0,1\}^K$ un masque indiquant la présence effective des commentaires. Un score d'attention est alors calculé pour chaque commentaire selon l'équation (3.1) :

$$e_k = \mathbf{v}^\top \tanh\left(\mathbf{W}\mathbf{h}_k\right), \tag{3.1}$$

où $\mathbf{W} \in \mathbb{R}^{d \times d}$ et $\mathbf{v} \in \mathbb{R}^d$ sont des paramètres appris.

Les poids d'attention normalisés sont ensuite définis à partir de l'équation (3.2) :

$$\alpha_k = \frac{\exp(e_k)\, m_k}{\sum_{j=1}^{K} \exp(e_j)\, m_j}, \tag{3.2}$$

ce qui garantit que seuls les commentaires valides contribuent à l'agrégation.

Enfin, la représentation globale des commentaires est obtenue par une somme pondérée, telle que définie dans l'équation (3.3) :

$$\mathbf{c} = \sum_{k=1}^{K} \alpha_k \mathbf{h}_k, \tag{3.3}$$

où $\mathbf{c} \in \mathbb{R}^d$ constitue l'embedding agrégé des commentaires associé à la vidéo.

3.4.4 PONDÉRATION DYNAMIQUE PAR MODALITÉ

Toutes les modalités ne contribuent pas de manière égale à la détection de contenus inappropriés. Certaines vidéos peuvent être visuellement neutres mais problématiques via l'audio ou les commentaires, tandis que d'autres contiennent des indices explicites directement

dans la modalité visuelle. Cette variabilité souligne le caractère fortement contextuel de la tâche. Afin de prendre en compte cette hétérogénéité, nous introduisons un mécanisme de *pondération dynamique* par modalité. L'objectif est d'adapter automatiquement l'importance relative de chaque source d'information en fonction du contenu analysé, plutôt que de supposer une contribution fixe.

Plus précisément, pour chaque représentation projetée, un réseau neuronal léger prédit un coefficient scalaire compris entre 0 et 1 via une fonction sigmoïde, permettant de moduler l'influence de la modalité correspondante.

Soit $\mathbf{z}_m \in \mathbb{R}^d$ la représentation projetée associée à la modalité $m \in \{\text{vid}, \text{aud}, \text{txt}, \text{comm}, \text{meta}\}$. Le coefficient de pondération est défini comme indiqué dans l'équation (3.4) :

$$g_m = \sigma\left(f_m(\mathbf{z}_m)\right), \tag{3.4}$$

où $f_m(\cdot)$ désigne un perceptron multicouches spécifique à la modalité m, et $\sigma(\cdot)$ correspond à la fonction sigmoïde.

La représentation modulée est ensuite obtenue conformément à l'équation (3.5) :

$$\tilde{\mathbf{z}}_m = g_m \cdot \mathbf{z}_m. \tag{3.5}$$

Ce mécanisme permet d'ajuster dynamiquement la contribution de chaque modalité en atténuant l'influence de celles qui sont peu informatives et en renforçant celles qui apportent des signaux discriminants. Il contribue ainsi à améliorer la robustesse du modèle face au bruit, aux données incomplètes et aux variations de contexte. La pondération dynamique est apprise conjointement avec le reste du modèle, sans supervision explicite sur l'importance relative des modalités, ce qui permet au système d'apprendre automatiquement les configurations les plus pertinentes.

3.4.5 MODÉLISATION DES INTERACTIONS PAR ATTENTION CROISÉE

Au-delà de la pondération individuelle des modalités, la détection de contenus inappropriés repose souvent sur des interactions complexes entre sources hétérogènes. Par exemple, une vidéo visuellement neutre peut devenir problématique lorsqu'elle est associée à un signal audio explicite ou à des commentaires alarmants. Ces observations soulignent que la nocivité d'un contenu ne réside pas uniquement dans une modalité isolée, mais émerge souvent de leur interaction. Il est donc essentiel de modéliser explicitement les dépendances inter-modales.

Dans cette optique, nous intégrons un mécanisme d'*attention croisée* multi-têtes, inspiré des travaux récents en fusion multimodale, et en particulier de l'architecture HICCAP (Baharlouei *et al.*, 2024). Dans HICCAP, l'attention croisée est principalement utilisée comme un mécanisme d'alignement sémantique entre modalités. À l'inverse, notre approche l'exploite comme un mécanisme d'interaction enrichie, permettant à chaque modalité d'intégrer dynamiquement des informations contextuelles issues des autres. Cette distinction est particulièrement pertinente dans le contexte de TikTok, où les contenus problématiques résultent fréquemment de la combinaison de plusieurs signaux.

Chaque modalité est représentée par un unique jeton vectoriel, obtenu après projection dans l'espace latent commun et application du mécanisme de pondération dynamique. Les interactions inter-modales sont ensuite modélisées via des blocs d'attention croisée multi-têtes opérant sur ces représentations.

Afin de limiter la complexité computationnelle et d'éviter des interactions redondantes, l'attention croisée est appliquée selon un schéma parcimonieux entre paires de modalités jugées sémantiquement pertinentes (par exemple texte↔audio, texte↔vidéo, audio↔vidéo), ainsi que pour l'enrichissement du texte par les commentaires et les métadonnées. Ce choix

permet de cibler les interactions les plus informatives tout en conservant une architecture efficace et stable.

Formellement, soit $\mathbf{q}_m \in \mathbb{R}^d$ le token associé à une modalité requête m, et $\mathbf{z}_n \in \mathbb{R}^d$ celui associé à une modalité source n. L'interaction inter-modale est modélisée à l'aide d'un module d'attention multi-têtes, noté $\mathrm{MHA}(\cdot)$, qui produit une représentation conditionnée telle que définie dans l'équation (3.6) :

$$\mathbf{h}_m = \mathrm{LayerNorm}\left(\mathbf{q}_m + \mathrm{MHA}(\mathbf{q}_m, \mathbf{z}_n)\right). \tag{3.6}$$

Dans sa formulation multi-têtes, le mécanisme d'attention est appliqué en parallèle sur plusieurs sous-espaces de représentation. Les différentes têtes sont ensuite concaténées puis projetées dans l'espace d'origine, comme indiqué dans l'équation (3.7) :

$$\mathrm{MHA}(\mathbf{q}_m, \mathbf{k}_n, \mathbf{v}_n) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_H)\mathbf{W}^O. \tag{3.7}$$

Ce mécanisme permet à chaque modalité d'intégrer de l'information provenant des autres, en tenant compte des dépendances contextuelles entre sources hétérogènes. L'utilisation d'une connexion résiduelle combinée à une normalisation de type *LayerNorm* contribue à stabiliser l'apprentissage et à préserver l'information initiale.

Les représentations ainsi enrichies intègrent des informations contextuelles issues des autres modalités et constituent les entrées de l'étape d'agrégation multimodale suivante.

3.4.6 AGRÉGATION SÉQUENTIELLE PAR BILSTM

Après enrichissement inter-modal, les représentations associées à chaque modalité sont concaténées pour former une séquence de longueur fixe, chaque position correspondant à une modalité. Cette séquence est ensuite traitée par un BiLSTM bidirectionnel, permettant de capturer des dépendances globales entre modalités.

La sortie du BiLSTM est ensuite agrégée à l'aide d'un mécanisme d'agrégation attentionnelle, produisant un vecteur global représentant la vidéo.

Le choix d'un BiLSTM est motivé par plusieurs considérations. Tout d'abord, la séquence à modéliser reste de faible longueur, correspondant à un nombre limité de modalités, ce qui rend ce type d'architecture particulièrement adapté. Ensuite, le BiLSTM permet de capturer des dépendances bidirectionnelles entre les différentes modalités, en tenant compte à la fois des interactions précédentes et suivantes dans la séquence. Enfin, il présente une complexité maîtrisée dans ce contexte, offrant un compromis efficace entre expressivité du modèle et coût computationnel.

Le vecteur global issu du module de fusion est ensuite transmis à une tête de classification basée sur un MLP, produisant les logits associés aux classes du dataset TikHarm. L'architecture est entraînée de bout en bout (à l'exception des parties gelées des encodeurs pré-entraînés), ce qui permet une intégration cohérente des différentes modalités tout en restant compatible avec des contraintes computationnelles réalistes.

Dans ce chapitre nous avons présenté la méthodologie adoptée, depuis la préparation et l'enrichissement du dataset jusqu'à la conception de l'architecture multimodale proposée. Les différentes étapes de prétraitement et d'extraction permettent de transformer des données

hétérogènes en représentations exploitables, tandis que le module de fusion cherche à modéliser les interactions entre les différentes modalités.

L'ensemble de ces choix méthodologiques constitue la base des expérimentations menées dans ce travail. Le Chapitre 4 décrit le protocole expérimental mis en place, en précisant les stratégies d'entraînement, d'optimisation ainsi que les métriques d'évaluation utilisées.

CHAPITRE IV
PROTOCOLE EXPÉRIMENTAL

Ce chapitre décrit le cadre expérimental mis en place pour l'évaluation de l'approche multimodale proposée. Il présente l'environnement matériel et logiciel utilisé, la configuration d'entraînement, les métriques d'évaluation retenues ainsi que les méthodes de comparaison issues de l'état de l'art et leur protocole de reproduction. L'objectif est de garantir la reproductibilité des expériences et la rigueur de la comparaison avec les baselines.

4.1 STRATÉGIE DE VALIDATION

Comme mentionné dans la section 3.1, le dataset enrichi est d'abord divisé en deux sous-ensembles disjoints : un ensemble d'apprentissage (90%) et un ensemble de test (10%), selon une partition stratifiée préservant la distribution des classes. Afin de réduire la variance liée à une partition unique et d'obtenir une estimation plus robuste des performances, une validation croisée stratifiée à cinq plis (5-fold cross-validation) est appliquée exclusivement sur l'ensemble d'apprentissage. À chaque itération, quatre plis (80%) sont utilisés pour l'entraînement et un pli (20%) pour la validation.

Les hyperparamètres du module de fusion, optimisés préalablement via Optuna sur une partition interne du jeu d'apprentissage, sont ensuite fixés afin d'assurer une comparaison cohérente entre les différents modèles et baselines. Un mécanisme d'early stopping basé sur l'exactitude de validation est appliqué dans chaque pli. Le nombre médian d'epochs retenu à travers les cinq plis est ensuite utilisé pour réentraîner le modèle sur l'intégralité de l'ensemble d'apprentissage. L'évaluation finale est réalisée une seule fois sur le jeu de test indépendant, garantissant l'absence de fuite d'information et une estimation non biaisée des performances.

4.2 ENVIRONNEMENT EXPÉRIMENTAL

L'ensemble des expériences a été implémenté en PyTorch et exécuté sur la plateforme Google Colab, en utilisant un Processeur graphique (Graphics Processing Unit) (GPU) NVIDIA A100-SXM4 disposant de 80 Go de mémoire. Cette configuration permet de gérer efficacement des modèles multimodaux intégrant des embeddings de grande dimension, tout en offrant un environnement reproductible. Le système utilisé repose sur Python 3.10. Les principales bibliothèques employées incluent PyTorch, Hugging Face Transformers, NumPy, scikit-learn et Optuna pour l'optimisation des hyperparamètres. Dans l'environnement expérimental considéré, les features multimodales pré-extraites (vidéo, audio, texte et métadonnées) sont préchargées en (Mémoire vive (Random Access Memory) (RAM)) avant l'entraînement. Cette stratégie réduit les accès disque au cours des epochs, au prix d'une consommation mémoire plus élevée, pouvant constituer une contrainte dans des environnements disposant de ressources RAM limitées.

4.3 CONFIGURATION D'ENTRAÎNEMENT

L'optimisation est réalisée à l'aide de l'optimiseur AdamW (Loshchilov & Hutter, 2017), avec une politique de taux d'apprentissage comprenant une phase de warm-up, suivie d'un cosine annealing. Afin de traiter le déséquilibre entre les classes, une pondération calculée à partir de la distribution du jeu d'entraînement est intégrée dans la fonction de perte, basée sur une entropie croisée avec label smoothing. La Précision mixte automatique (Automatic Mixed Precision) (AMP) est activée afin d'accélérer l'entraînement et de réduire la consommation mémoire GPU. Un mécanisme Moyenne mobile exponentielle (Exponential Moving Average) (EMA) des poids est également utilisé pour stabiliser l'optimisation.

4.3.1 OPTIMISATION DES HYPERPARAMÈTRES

Les hyperparamètres du module de fusion multimodale sont sélectionnés à l'aide d'une optimisation bayésienne via la bibliothèque Optuna. L'espace de recherche inclut notamment la dimension latente commune, le taux de dropout, le taux d'apprentissage, les termes de régularisation et les paramètres liés aux mécanismes de fusion. La recherche est menée sur un ensemble de 30 essais. La meilleure configuration retenue est présentée dans le Tableau 4.1.

TABLEAU 4.1 : Hyperparamètres du modèle

Hyperparamètre	**Valeur**
Dimension latente commune (d_{fuse})	1024
Nombre de têtes d'attention (n_{heads})	8
Taux de dropout	0.1008
Taux d'apprentissage initial	1.48×10^{-4}
Weight decay	2.08×10^{-5}
Coefficient de label smoothing	0.0342
Probabilité de modality dropout (p_{moddrop})	0.0710
Désactivation de la régularisation L_2 sur les couches de projection	Oui
Terme de régularisation des mécanismes de gating	7.99×10^{-4}

Cette configuration est utilisée pour l'ensemble des expériences rapportées dans la suite du mémoire.

4.4 MÉTRIQUES D'ÉVALUATION

Les performances des modèles sont évaluées sur le jeu de test à l'aide de plusieurs métriques complémentaires. L'exactitude est d'abord utilisée afin de mesurer la proportion globale de prédictions correctes. Des métriques plus fines, telles que la précision, le rappel et

le F1-score, sont également considérées, sous une forme macro-pondérée afin de prendre en compte l'ensemble des classes de manière équilibrée. Par ailleurs, le coefficient de corrélation de Matthews (Coefficient de corrélation de Matthews (MCC)) est utilisé en raison de sa pertinence dans des contextes potentiellement déséquilibrés. Enfin, l'aire sous la courbe ROC (Caractéristique de fonctionnement du récepteur (Receiver Operating Characteristic) (ROC)-Aire sous la courbe (Area Under the Curve) (AUC)) permet d'évaluer la capacité du modèle à discriminer efficacement entre les différentes classes. Nous reportons également le nombre d'epochs nécessaires à la convergence, ainsi que le temps d'entraînement, afin d'analyser le coût computationnel des différentes approches.

4.5 MÉTHODES DE COMPARAISON (BASELINES)

Pour situer les performances de notre approche, nous comparons notre modèle à plusieurs méthodes représentatives de l'état de l'art, couvrant à la fois des approches unimodales et multimodales proposées pour la détection de contenus inappropriés sur YouTube ou TikTok. Les méthodes considérées incluent notamment Samba (Binh *et al.*, 2022), Enhanced Multimodal Content Moderation of Children's Videos using Audiovisual Fusion (Ahmed *et al.*, 2024), HICCAP (Baharlouei *et al.*, 2024), mTIKGUARD (Nguyen *et al.*, 2025) ainsi que Leveraging Multi-Modality and Enhanced Temporal Networks for Robust Violence Detection (Na *et al.*, 2024). Ces méthodes ont été sélectionnées en raison de leur pertinence pour la tâche étudiée et de la diversité des modalités exploitées.

4.6 REPRODUCTION ET ADAPTATION DES BASELINES

Afin d'assurer une comparaison cohérente, toutes les baselines sont évaluées sur la même partition du dataset TikHarm enrichi, avec un prétraitement cohérent des données.

Samba (Binh *et al.*, 2022) : Le code officiel étant disponible, il a été réutilisé directement. La tête de classification a été adaptée pour produire quatre sorties correspondant aux classes du dataset TikHarm. En l'absence de transcriptions automatiques, les descriptions et commentaires TikTok ont été concaténés afin de constituer une entrée textuelle équivalente. Les miniatures ont été obtenues en extrayant un frame par vidéo.

Enhanced Multimodal Content Moderation of Children's Videos using Audiovisual Fusion (Ahmed *et al.*, 2024) : Le code source n'étant pas public, l'architecture a été reproduite à partir des descriptions méthodologiques et des hyperparamètres rapportés. Les principales adaptations incluent le passage d'une classification binaire à quatre classes, l'ajustement des prompts textuels au contexte TikTok et la ré-implémentation d'une perte contrastive de type CLIP combinée à une entropie croisée supervisée.

HICCAP (Baharlouei *et al.*, 2024) : Le code officiel étant disponible et reposant sur des technologies similaires à celles de notre approche (I3D, VGGish, BERT), les features extraites ont été directement réutilisées, et la tête de classification a été adaptée pour prédire les classes du dataset TikHarm.

Leveraging Multi-Modality and Enhanced Temporal Networks for Robust Violence Detection (Na *et al.*, 2024) : Le code officiel n'étant pas disponible, l'architecture a été reproduite à partir des descriptions fournies. Les captions générées automatiquement dans le cadre original ont été remplacées par les descriptions et commentaires TikTok concaténés.

mTIKGUARD (Nguyen *et al.*, 2025) : Le code officiel étant disponible, il a été réutilisé afin de garantir une reproduction fidèle de l'architecture proposée. Cette méthode s'appuie

également sur une version enrichie du dataset TikHarm comportant quatre classes, ce qui ne nécessite pas d'adaptation de la tête de classification. Le modèle est entraîné et évalué sur notre partition du dataset enrichi, avec un prétraitement cohérent afin d'assurer une comparaison équitable.

Malgré les efforts déployés pour reproduire fidèlement ces méthodes, l'absence de code officiel pour certaines baselines peut introduire des écarts d'implémentation. Les résultats doivent donc être interprétés comme une comparaison expérimentale menée dans un cadre contrôlé, plutôt que comme une réplication exacte.

Ce protocole expérimental a été conçu pour assurer une évaluation rigoureuse et homogène de l'approche proposée. En s'appuyant sur un cadre d'entraînement commun, des métriques complémentaires et des baselines adaptées, il permet d'analyser de manière fiable l'apport de la multimodalité ainsi que des mécanismes de fusion introduits. Le Chapitre 5 présente les résultats expérimentaux obtenus et en propose une analyse détaillée.

CHAPITRE V
RÉSULTATS ET ANALYSES

Ce chapitre présente les résultats expérimentaux obtenus par l'approche multimodale proposée et les compare aux différentes baselines issues de l'état de l'état. Il propose ensuite une analyse des écarts de performance observés, avant d'examiner plus en détail l'apport des différentes modalités ainsi que l'impact des choix architecturaux à travers une étude d'ablation et une analyse qualitative.

5.1 RÉSULTATS

Dans un premier temps, nous évaluons les performances globales des différentes approches sur le jeu de test du dataset TikHarm enrichi, afin de comparer leur capacité à détecter des contenus inappropriés dans un cadre multimodal. L'évaluation repose sur plusieurs métriques complémentaires, incluant l'accuracy, la précision, le rappel, le F1-score, le coefficient de corrélation de Matthews (MCC) ainsi que l'aire sous la courbe ROC (ROC-AUC).

Les résultats obtenus sont synthétisés dans le Tableau 5.1. Ils montrent que le modèle multimodal proposé surpasse l'ensemble des baselines sur toutes les métriques considérées. Il atteint notamment une accuracy de 87.13%, un F1-score de 87.14% et un MCC de 82.96%, ce qui traduit à la fois une bonne capacité de classification globale et une robustesse face au déséquilibre des classes.

TABLEAU 5.1 : Comparaison des performances sur les données de test TikHarm

Approches	Acc (%)	Précision (%)	Rappel (%)	F1-score (%)	MCC (%)	ROC-AUC (%)	Epoch	Time
HICCAP Baharlouei *et al.* (2024)	84.50	84.80	84.77	84.43	79.57	96.55	15	9.25 min
SAMBA Binh *et al.* (2022)	71.05	71.58	71.38	71.23	61.48	89.19	38	0.39 min
Ahmed *et al.* (2024)	81.87	82.19	81.74	81.85	75.85	96.59	4	5.69 min
Na *et al.* (2024)	62.57	68.79	63.27	63.11	51.83	85.65	19	31.25 min
mTIKGUARD Nguyen *et al.* (2025)	83.92	84.13	84.23	83.91	78.74	96.02	5	284.86 min
SoSAFE (ours)	**87.13**	**87.19**	**87.43**	**87.14**	**82.96**	**97.06**	**20**	**5.1 min**

Comparé aux méthodes de référence reproduites dans le même cadre expérimental, notre modèle surpasse la meilleure baseline concurrente Baharlouei *et al.* (2024) de 2.6 points d'accuracy et de 2,7 points de F1-score. Malgré une architecture multimodale plus riche, il maintient également un coût computationnel raisonnable, avec un temps d'entraînement inférieur à la majorité des baselines.

5.2 ÉTUDE D'ABLATION

Afin de mieux comprendre l'impact des différents choix méthodologiques et architecturaux, nous avons mené une étude d'ablation systématique. Cette analyse vise à évaluer l'influence de plusieurs facteurs, notamment la stratégie de partition des données, les modalités exploitées, la structure de fusion multimodale, ainsi que les mécanismes de pondération et d'interaction entre modalités.

Pour des raisons de coût computationnel et afin de garantir la comparabilité entre les différentes configurations, les expériences d'ablation ont été réalisées sur une partition fixe des données (train/validation/test), plutôt que dans le cadre du protocole de validation croisée en k-fold utilisé pour l'évaluation finale du modèle. Cette stratégie permet d'isoler l'effet de chaque composant architectural tout en maintenant un cadre expérimental stable.

Les résultats correspondants sont présentés dans le Tableau 5.2, qui synthétise l'impact des différentes configurations sur les performances du modèle.

TABLEAU 5.2 : Étude d'ablation de l'architecture multimodale proposée sur le dataset TikHarm (jeu de test)

Configuration	Modalities	Acc (%)	F1 (%)	Prec (%)	Rec (%)	MCC (%)
Modèle final (Cross Validation (CV))	**Vidéo (V) + Audio (A) + Description (D) + Commentaires (C) +Métadonnées (M)**	**87.13**	**87.14**	**87.19**	**87.43**	**82.96**
Partition (70/15/15)	V + A + D + C + M	86.00	86.00	87.00	86.00	81.00
Sans métadonnées	V + A + D + C	86.50	86.50	86.50	86.80	82.00
Sans commentaires	V + A + D + M	85.00	85.10	85.10	85.20	80.00
Fusion Transformer (sans BiLSTM)	V + A + D + C + M	86.50	86.50	86.60	86.80	82.10
Optimisation explicite des poids modaux	V + A + D + C + M	86.20	86.20	86.30	86.60	81.80
Optimisation greedy shapley	V + A + D + C + M	82.70	82.70	82.70	82.90	77.00

5.2.1 IMPACT DE LA STRATÉGIE DE VALIDATION

Nous avons initialement évalué l'architecture proposée en utilisant une partition 70/15/15 (entraînement/validation/test), conformément à certains travaux antérieurs. Cette configuration conduit à une légère dégradation des performances, avec une accuracy et un F1-score d'environ 86%, par rapport à la configuration finale adoptée. Ce résultat souligne l'importance d'un compromis adéquat entre quantité de données d'entraînement et fiabilité de l'évaluation, en particulier dans un contexte multimodal où les modèles sont fortement paramétrés.

5.2.2 CONTRIBUTION DES MODALITÉS SOCIALES : COMMENTAIRES ET MÉTADONNÉES

L'ablation des commentaires et des métadonnées permet d'évaluer leur apport spécifique. Lorsque les commentaires sont retirés, le F1-score chute à 85.1 %, tandis que la suppression des métadonnées entraîne une baisse plus modérée à 86.5%. Ces résultats confirment que, bien que secondaires par rapport aux signaux audiovisuels, les modalités sociales fournissent un

contexte interprétatif précieux, notamment pour la désambiguïsation de contenus implicites ou visuellement neutres.

5.2.3 ANALYSE DE LA STRUCTURE DE FUSION

Nous avons également comparé différentes stratégies de fusion multimodale. Le remplacement du BiLSTM bidirectionnel par un Transformer à deux couches conduit à un F1-score de 86.5 %, inférieur à celui obtenu avec le BiLSTM. Ce résultat suggère que, dans un cadre où la séquence est courte et correspond à un nombre limité de modalités, le BiLSTM constitue un choix plus adapté et plus stable que des architectures auto-attentives plus complexes.

5.2.4 ANALYSE DE L'OPTIMISATION GREEDY SHAPLEY

Nous avons également évalué une stratégie d'optimisation explicite des poids modaux fondée sur une approximation greedy des valeurs de Shapley. Cette analyse a été réalisée a posteriori sur le meilleur modèle sauvegardé lors de l'entraînement, sélectionné sur la base de ses performances sur le jeu de validation. Les représentations multimodales et les paramètres du modèle ont ainsi été fixés, et seules les contributions relatives des modalités ont été estimées.

Cette approche vise à quantifier la contribution marginale de chaque modalité et à en déduire des poids globaux optimaux pour la fusion. Toutefois, les résultats montrent que cette stratégie conduit à une dégradation notable des performances, avec une accuracy de 82.7 % et un F1-score de 82.7 %, soit une baisse de près de 4,5 points de F1 par rapport à la fusion multimodale complète proposée. Une diminution similaire est observée sur le MCC, qui chute à 77.00

Ce comportement peut s'expliquer par plusieurs facteurs. Tout d'abord, l'optimisation greedy Shapley repose sur des poids statiques, identiques pour l'ensemble des vidéos, alors que la pertinence relative des modalités varie fortement selon le contenu analysé. Sur TikTok, une modalité peu informative pour une vidéo donnée peut devenir déterminante dans un autre contexte, ce que des poids globaux ne permettent pas de capturer.

Par ailleurs, cette approche n'intègre pas explicitement les interactions inter-modales, puisque les contributions sont estimées de manière additive. Or, nos résultats montrent que les signaux nuisibles émergent fréquemment de la combinaison et de la cohérence (ou incohérence) entre plusieurs modalités, par exemple entre le contenu visuel et les réactions des utilisateurs.

Ces résultats confirment que, dans le contexte de TikTok, des mécanismes dynamiques et conditionnels, tels que le gating par modalité et la cross-attention inter-modale, sont mieux adaptés que des stratégies d'optimisation explicite basées sur des contributions moyennes. Les valeurs de Shapley apparaissent ainsi plus pertinentes comme outil d'analyse et d'interprétabilité, permettant d'expliquer a posteriori les décisions du modèle, plutôt que comme mécanisme direct de fusion ou d'optimisation des poids.

5.2.5 PONDÉRATION EXPLICITE DES MODALITÉS

Enfin, nous avons étudié l'impact de l'optimisation explicite des poids des modalités à l'aide d'Optuna. Cette stratégie n'a pas permis d'améliorer les performances (F1-score de 86.2 %), ce qui suggère que les mécanismes de gating dynamique et de cross-attention intégrés dans l'architecture sont déjà suffisants pour apprendre efficacement l'importance relative des différentes modalités. Dans l'ensemble, cette étude d'ablation montre que :

— la partition des données influence de manière non négligeable les performances,

— les commentaires et métadonnées apportent un gain mesurable,

— le BiLSTM est mieux adapté que le Transformer pour l'agrégation finale,

— une cross-attention exhaustive n'est pas nécessaire,

— et les stratégies de fusion naïves sont largement insuffisantes.

Ces résultats confirment la pertinence des choix architecturaux retenus pour la détection multimodale de contenus inappropriés sur TikTok.

Ils nécessitent toutefois une analyse approfondie afin d'en comprendre les implications, les limites et les apports. Dans cette perspective, le chapitre 6 propose une discussion détaillée de ces résultats.

CHAPITRE VI
DISCUSSION

Ce chapitre analyse les résultats obtenus par l'approche multimodale proposée et les met en perspective avec les choix méthodologiques effectués. Il examine le rôle des différentes modalités, le comportement du modèle selon les classes du dataset TikHarm, ainsi que les implications méthodologiques et pratiques pour la modération automatique de contenus destinés aux enfants.

6.1 DISCUSSION DES RÉSULTATS

Les performances obtenues confirment la pertinence d'une approche multimodale complète pour la détection de contenus inappropriés sur TikTok. Elles apportent également des éléments de réponse aux questions de recherche formulées dans ce travail. Plus précisément, elles permettent de mieux comprendre dans quelle mesure l'intégration conjointe de signaux audiovisuels, textuels et comportementaux améliore la détection de contenus inappropriés, tout en mettant en évidence l'apport des mécanismes de fusion adaptative dans la modélisation des interactions entre modalités.

La première hypothèse supposait qu'une architecture intégrant simultanément des signaux visuels, audio, textuels ainsi que des métadonnées d'engagement permettrait d'améliorer la détection de contenus inappropriés par rapport aux approches unimodales ou multimodales partielles. Les résultats expérimentaux confirment clairement cette hypothèse : le modèle proposé obtient des performances supérieures aux baselines reproduites dans le même cadre expérimental, ce qui souligne l'intérêt de combiner plusieurs sources d'information complémentaires. Cette amélioration peut s'expliquer par la diversité des formes que peuvent

prendre les contenus problématiques sur TikTok : certains reposent sur des indices visuels explicites, tandis que d'autres sont davantage suggérés par l'audio, le texte ou les dynamiques d'engagement associées. Cette complémentarité des modalités apparaît ainsi comme un facteur déterminant pour une détection robuste.

L'étude d'ablation met également en évidence l'importance des modalités textuelles et des métadonnées d'engagement. La suppression de ces informations entraîne une baisse mesurable des performances, ce qui confirme leur rôle dans la contextualisation du contenu. Les métadonnées fournissent notamment des indices indirects sur la diffusion et la réception des vidéos, ce qui s'avère particulièrement utile pour détecter des contenus implicitement problématiques, c'est-à-dire des contenus dont la nocivité ne peut être déduite à partir d'une seule modalité.

La seconde hypothèse portait sur l'apport des mécanismes de fusion adaptative, tels que le mécanisme de pondération dynamique par modalité et l'attention croisée inter-modale. Les résultats obtenus, notamment à travers les expériences d'ablation, suggèrent que ces mécanismes permettent de mieux modéliser les interactions entre modalités que des approches de fusion plus simples, telles que la concaténation ou les pondérations fixes. En particulier, la capacité du modèle à adapter dynamiquement l'importance des différentes sources d'information contribue à améliorer sa robustesse face à des signaux bruités, ambigus ou partiellement informatifs. Cette adaptation contextuelle apparaît comme un élément clé dans un environnement où la pertinence des modalités varie fortement d'une vidéo à l'autre.

Par ailleurs, la comparaison avec les approches existantes met en évidence que les architectures initialement développées pour des plateformes de vidéos longues ne se transfèrent pas directement au contexte TikTok. Les méthodes centrées principalement sur les interactions audiovisuelles restent performantes pour des contenus explicitement perturbants, mais montrent

leurs limites lorsque les signaux nuisibles sont implicites, ambigus ou distribués entre plusieurs modalités. L'amélioration observée avec notre modèle suggère ainsi que l'intégration explicite d'informations contextuelles et comportementales constitue un facteur différenciant dans ce contexte, en permettant de mieux capturer la nature socio-contextuelle des contenus.

Dans l'ensemble, ces résultats confirment que la détection de contenus inappropriés sur TikTok ne repose pas uniquement sur la présence d'indices visuels explicites, mais sur l'interaction entre plusieurs modalités et leur contextualisation. Elle ne peut être réduite à une simple tâche de classification mais relève d'un processus décisionnel complexe, dans lequel la nocivité émerge de l'interaction entre des signaux audiovisuels et des indices sociaux contextualisés. Ils soulignent également l'importance de concevoir des architectures capables de modéliser explicitement ces interactions afin de mieux refléter la complexité des contenus diffusés sur les plateformes de vidéos courtes.

6.1.1 RÔLE DES DIFFÉRENTES MODALITÉS

Les descriptions constituent un signal intentionnel produit par le créateur du contenu. Elles jouent un rôle important dans la contextualisation de vidéos ambiguës, notamment en permettant de distinguer un message de prévention d'un contenu potentiellement incitatif. Toutefois, elles peuvent également être stratégiques, ironiques ou partiellement déconnectées du contenu réel. Leur apport doit donc être interprété avec prudence et réside principalement dans leur complémentarité avec les autres modalités.

Les commentaires apportent une dimension collective essentielle. Ils reflètent la manière dont le contenu est perçu et interprété par la communauté, et peuvent contenir des avertissements explicites, des réactions émotionnelles ou des formes d'ironie révélatrices. Bien qu'ils soient souvent bruités et hétérogènes, ils contribuent à réduire certaines erreurs de

classification, en particulier pour des contenus implicitement problématiques dont la nocivité dépend du contexte.

Les métadonnées d'engagement fournissent un signal comportemental indirect lié aux dynamiques de diffusion et de réception du contenu. Si leur contribution individuelle reste plus modeste, leur intégration permet d'améliorer la robustesse globale du modèle lorsqu'elles sont combinées aux autres modalités. Elles apportent ainsi une information complémentaire sur la manière dont le contenu circule et est interprété au sein de la plateforme.

L'apport différencié des modalités souligne que la détection repose sur une intégration contextuelle plutôt que sur un signal dominant unique, renforçant l'idée d'une modération fondamentalement multimodale et socio-contextuelle.

6.1.2 ANALYSE PAR CLASSE

L'analyse des performances par classe montre que les erreurs concernent majoritairement des catégories conceptuellement proches.

La classe *Adult Content* est détectée avec une précision et un rappel élevés, indiquant que les signaux explicites sont correctement identifiés. Les confusions observées avec *Harmful Content* reflètent la proximité sémantique entre certains contenus sexuellement explicites et des comportements jugés dangereux.

La classe *Harmful Content* présente des performances légèrement inférieures, en raison de son hétérogénéité. Elle regroupe des comportements à risque ne présentant pas toujours de marqueurs visuels explicites, ce qui complique la classification.

La classe *Safe* montre des performances équilibrées, traduisant une capacité satisfaisante à limiter les faux positifs, un enjeu crucial pour éviter la surcensure.

Enfin, la classe *Suicide* est globalement bien reconnue. Les confusions observées concernent principalement *Harmful Content*, ce qui s'explique par la proximité entre comportements à risque et références implicites à des thématiques suicidaires.

De manière générale, les erreurs reflètent davantage la complexité conceptuelle des catégories que des faiblesses structurelles du modèle.

6.1.3 DISCUSSION ARCHITECTURALE

Les résultats confirment que l'exploitation conjointe des modalités visuelle, audio, textuelle et comportementale constitue un facteur déterminant de performance. Les gains observés dans l'étude d'ablation montrent que les modalités sociales apportent une information complémentaire non redondante.

Le mécanisme de gating par modalité permet d'ajuster dynamiquement l'importance relative de chaque source d'information. Cette adaptation contextuelle s'avère plus efficace qu'une pondération fixe ou globale, particulièrement dans un environnement où la pertinence des modalités varie fortement d'une vidéo à l'autre.

La cross-attention inter-modale contribue à modéliser explicitement les interactions entre modalités. Toutefois, une cross-attention exhaustive n'apporte pas de gain significatif par rapport à une stratégie ciblée, ce qui suggère que des interactions pertinentes et limitées sont suffisantes dans un cadre à nombre restreint de modalités.

Le choix d'un BiLSTM bidirectionnel pour l'agrégation finale se révèle adapté à la courte séquence modale considérée. Contrairement aux architectures Transformer, conçues pour de longues séquences, le BiLSTM offre ici un compromis efficace entre expressivité et stabilité.

6.1.4 DISCUSSION SUR L'OPTIMISATION EXPLICITE DES POIDS MODAUX

Les stratégies de pondération explicite des modalités, qu'elles reposent sur une recherche bayésienne ou sur une analyse de type Shapley, n'améliorent pas les performances.

Dans notre cadre expérimental, l'analyse de type Shapley est utilisée à posteriori comme outil d'interprétabilité. Lorsque ces contributions moyennes sont utilisées comme contraintes explicites de fusion, les performances se dégradent. Cela indique que l'importance des modalités ne peut être capturée par des poids statiques globaux.

À l'inverse, les mécanismes dynamiques intégrés dans l'architecture permettent d'adapter la fusion au contexte spécifique de chaque vidéo. Les méthodes de type Shapley apparaissent ainsi plus pertinentes pour l'analyse explicative que pour l'optimisation directe.

Dans l'ensemble, ces analyses mettent en évidence les facteurs clés de performance du modèle proposé, tout en soulignant l'importance d'une modélisation multimodale adaptée au contexte des plateformes de vidéos courtes. Néanmoins, plusieurs limites doivent être considérées afin de nuancer la portée de ces résultats.

6.2 LIMITES DU TRAVAIL

Malgré les performances obtenues, plusieurs limites doivent être prises en compte afin de situer correctement la portée des résultats.

6.2.1 LIMITES LIÉES AU DATASET

Le dataset TikHarm, bien qu'il constitue l'un des rares jeux de données publics dédiés à la modération de contenus TikTok, reste de taille modérée au regard de la diversité réelle

des contenus présents sur la plateforme. Cette contrainte peut limiter la capacité du modèle à généraliser à des cas rares ou à des tendances émergentes.

Par ailleurs, la prédominance d'une langue et d'un contexte culturel spécifiques peut introduire un biais dans les représentations apprises. Une évaluation sur des données multilingues et culturellement diversifiées serait nécessaire pour confirmer la robustesse du modèle à grande échelle.

Enfin, certaines frontières entre classes, notamment entre *Harmful Content* et *Suicide*, demeurent conceptuellement floues. Cette ambiguïté intrinsèque se reflète à la fois dans l'annotation humaine et dans certaines erreurs de classification.

6.2.2 LIMITES LIÉES À L'ENRICHISSEMENT DES DONNÉES

L'enrichissement du dataset par l'ajout des descriptions, commentaires et métadonnées repose sur un processus de collecte automatisé dépendant de la disponibilité des contenus sur la plateforme. La suppression ou la modification de vidéos dans le temps peut affecter la reproductibilité complète des expériences.

De plus, les descriptions et commentaires peuvent être bruités, ironiques ou volontairement trompeurs. Bien que le cadre multimodal permette d'atténuer ces effets, ces phénomènes restent difficiles à modéliser parfaitement.

6.2.3 LIMITES MÉTHODOLOGIQUES

Le fine-tuning partiel des encodeurs pré-entraînés constitue un compromis entre performance et coût computationnel. Si cette stratégie réduit le risque de sur-apprentissage et accélère l'entraînement, elle peut également restreindre l'adaptation complète aux spécificités

du domaine TikTok. Plus généralement, l'utilisation d'encodeurs standard pré-entraînés peut ne pas permettre de saisir pleinement l'évolution des formes de nocivité spécifiques à chaque plateforme, dont les dynamiques culturelles et les modes d'expression évoluent rapidement.

Par ailleurs, bien que la stratégie de préchargement en mémoire vive des features avant l'entraînement permette de stabiliser et d'accélérer significativement l'entraînement, elle implique une consommation mémoire importante. Cette contrainte peut limiter la reproductibilité de l'approche dans des environnements disposant de ressources matérielles plus restreintes.

Enfin, bien que plusieurs baselines aient été reproduites dans un cadre expérimental homogène, certaines méthodes récentes n'ont pas pu être évaluées dans des conditions strictement identiques en raison des contraintes d'implémentation ou de l'absence de code public. Les comparaisons doivent donc être interprétées comme des évaluations expérimentales contrôlées plutôt que comme des réplications exactes. De plus, bien que des analyses d'interprétabilité aient été menées à des fins exploratoires, l'intégration de mécanismes explicites d'explication des décisions reste un défi ouvert.

Ce chapitre a permis d'analyser les résultats obtenus et de mieux comprendre le comportement du modèle proposé.

CONCLUSION ET PERSPECTIVES

Ce travail s'est inscrit dans la problématique de la détection automatique de contenus inappropriés destinés aux enfants sur la plateforme TikTok, un environnement caractérisé par la brièveté des vidéos, leur forte éditorialisation et l'importance des dynamiques sociales. Face aux limites des approches unimodales et des architectures multimodales génériques, nous avons proposé une architecture spécifiquement conçue pour exploiter la richesse multimodale et contextuelle des contenus TikTok.

L'approche SoSAFE développée repose sur l'intégration conjointe de cinq modalités complémentaires : la vidéo, l'audio, les descriptions, les commentaires et les métadonnées d'engagement. À travers un module de fusion combinant projection dans un espace latent commun, mécanismes de gating dynamique, cross-attention ciblée et agrégation séquentielle par BiLSTM, le modèle apprend à adapter la pondération des modalités en fonction du contexte de chaque vidéo et à modéliser explicitement leurs interactions.

Les résultats expérimentaux obtenus sur le dataset TikHarm enrichi montrent que cette approche permet d'améliorer de manière notable les performances par rapport aux baselines reproduites dans un cadre expérimental homogène, tout en conservant un coût computationnel maîtrisé. Les analyses menées, notamment l'étude d'ablation et l'évaluation par classe, confirment que si la modalité visuelle demeure centrale, les signaux sociaux, en particulier les commentaires et les métadonnées jouent un rôle déterminant dans la désambiguïsation de contenus implicites ou contextuellement sensibles.

Ces résultats soulignent qu'une modération efficace sur TikTok ne peut se limiter à l'analyse audiovisuelle isolée, mais doit intégrer la dimension interprétative et sociale propre aux plateformes de vidéos courtes.

Ils permettent également de répondre explicitement aux questions de recherche formulées. Concernant la première question (QR1), l'étude confirme que l'intégration des commentaires et des métadonnées d'engagement améliore les performances de détection en apportant un contexte social essentiel à l'interprétation de contenus implicites. Concernant la seconde question (QR2), les résultats montrent que les mécanismes de fusion adaptative, notamment le gating dynamique et la cross-attention inter-modale, surpassent les stratégies de fusion statiques en permettant une modélisation plus flexible et contextuelle des interactions entre modalités.

Plusieurs perspectives de recherche se dégagent de ce travail. Une première consiste à étendre l'évaluation à des jeux de données plus larges, multilingues et culturellement diversifiés, afin d'évaluer la robustesse du modèle dans des contextes variés. Une seconde piste concerne l'intégration de modèles multimodaux de nouvelle génération, notamment des modèles fondation vidéo-langage, tout en conservant des stratégies de fusion adaptées aux contraintes spécifiques des vidéos courtes.

L'amélioration de l'interprétabilité multimodale constitue également une direction importante, afin de mieux comprendre les interactions entre modalités et de renforcer la confiance dans les systèmes de modération automatique. Enfin, l'étude de stratégies hybrides combinant détection algorithmique et validation humaine apparaît comme une voie prometteuse pour un déploiement opérationnel à grande échelle.

En définitive, ce travail met en évidence que la détection de contenus inappropriés sur TikTok nécessite des architectures explicitement adaptées à sa nature multimodale et socio-contextuelle. Les résultats obtenus ouvrent la voie à des systèmes de modération plus robustes, plus adaptatifs et mieux alignés avec les enjeux de protection des jeunes utilisateurs.

BIBLIOGRAPHIE

Ahmed, S. H., Khan, M. J. & Sukthankar, G. (2024). Enhanced Multimodal Content Moderation of Children's Videos using Audiovisual Fusion. *arXiv preprint arXiv :2405.06128*.

Anderson, K. E. (2020). Getting acquainted with social networks and apps : it is time to talk about TikTok. *Library hi tech news*, *37*(4), 7–12.

Apify (2025). *TikTok Scraper*. https://apify.com/clockworks/tiktok-scraper.

Baharlouei, E., Shafaei, M., Zhang, Y., Escalante, H. J. & Solorio, T. (2024). Labeling Comic Mischief Content in Online Videos with a Multimodal Hierarchical-Cross-Attention Model. *arXiv preprint arXiv :2406.07841*.

Balat, M., Gabr, M., Bakr, H. & Zaky, A. B. (2024). TikGuard : A Deep Learning Transformer-Based Solution for Detecting Unsuitable TikTok Content for Kids. Dans *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pp. 337–340. IEEE.

Binh, L., Tandon, R., Oinar, C., Liu, J., Durairaj, U., Guo, J., Zahabizadeh, S., Ilango, S., Tang, J., Morstatter, F. *et al.* (2022). Samba : Identifying inappropriate videos for young children on YouTube. Dans *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 88–97.

Bonifazi, G., Cecchini, S., Corradini, E., Giuliani, L., Ursino, D. & Virgili, L. (2024). Investigating community evolutions in TikTok dangerous and non-dangerous challenges. *Journal of Information Science*, *50*(5), 1170–1194.

Carreira, J. & Zisserman, A. (2017). Quo vadis, action recognition ? a new model and the kinetics dataset. Dans *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308.

Chuttur, M. Y. & Nazurally, A. (2022). A multi-modal approach to detect inappropriate cartoon video contents using deep learning networks. *Multimedia Tools and Applications*, *81*(12), 16881–16900.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. Dans *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics : human language technologies, volume 1 (long and short papers)*, pp. 4171–4186.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. Dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B. *et al.* (2017). CNN architectures for large-scale audio classification. Dans *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135. IEEE.

Ishikawa, A., Bollis, E. & Avila, S. (2019). Combating the elsagate phenomenon : Deep learning architectures for disturbing cartoons. Dans *2019 7th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6. IEEE.

Khan, N. F., Amin, S. U., Jan, Z. & Yan, C. (2024). The Detection of Violent Scenes in Cartoon Movies Using Deep Learning Approach. *IEEE Access*.

Kriegel, E. R., Lazarevic, B., Athanasian, C. E. & Milanaik, R. L. (2021). TikTok, Tide Pods and Tiger King : health implications of trends taking over pediatric populations. *Current opinion in pediatrics*, *33*(1), 170–177.

López, B. C. (2021). Ambivalencia en TikTok : aprendizaje permanente y riesgos de seguridad coexistiendo. *IE Revista de Investigación Educativa de la REDIECH*, p. 68.

Loshchilov, I. & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv :1711.05101*.

Minhaj, F. & Leonard, J. (2021). *Dangers of the TikTok Benadryl Challenge*. Contemporary Pediatrics. Article en ligne, consulté le 29 juin 2026, URL : https ://www.contemporarypediatrics.com/view/dangers-of-the-tiktok-benadryl-challenge.

Na, G., Ko, J. & Cheoi, K. (2024). Leveraging Multi-Modality and Enhanced Temporal

Networks for Robust Violence Detection. *Machine Learning and Knowledge Extraction*, *6*(4), 2422–2434.

Nguyen, D. T., Lam, N. H., Nguyen, A. H.-T. & Do, T.-H. (2025). MTikGuard System : A Transformer-Based Multimodal System for Child-Safe Content Moderation on TikTok. *arXiv preprint arXiv :2511.17955*.

Singh, G. V., Ghosh, S., Firdaus, M., Ekbal, A. & Bhattacharyya, P. (2025). Unmasking offensive content : a multimodal approach with emotional understanding. *Multimedia Tools and Applications*, pp. 1–24.

Singh, S., Kaushal, R., Buduru, A. B. & Kumaraguru, P. (2019). KidsGUARD : fine grained approach for child unsafe video representation and detection. Dans *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*, pp. 2104–2111.

Slotta, D. (2025). *TikTok – statistiques et faits*. Statista. Publié le 17 décembre 2025., Repéré le 2025-12-28, à https://www.statista.com/topics/6077/tiktok/

Tahir, R., Ahmed, F., Saeed, H., Ali, S., Zaffar, F. & Wilson, C. (2019). Bringing the kid back into youtube kids : Detecting inappropriate content on video streaming platforms. Dans *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pp. 464–469.

Yousaf, K. & Nawaz, T. (2022). A deep learning-based approach for inappropriate content detection and classification of youtube videos. *IEEE Access*, *10*, 16283–16298.