UQAC
Université du Québec
à Chicoutimi

FUSION DYNAMIQUE TEXTE-ENGAGEMENT POUR LA DÉTECTION D'IDÉATION SUICIDAIRE SUR TIKTOK À PARTIR DE DONNÉES LIMITÉES

PAR ATHAR ESSAFI

MÉMOIRE

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

QUÉBEC, CANADA

RÉSUMÉ

La détection automatique de l'idéation suicidaire sur TikTok soulève des difficultés particulières que les approches conçues pour les plateformes textuelles classiques ne traitent pas. Les créateurs remplacent souvent les mots sensibles par des codes destinés à contourner la modération, un phénomène linguistique appelé algospeak. De plus, le texte associé à une vidéo est fragmenté entre le titre, les mots-clics, la transcription audio et les commentaires de l'auditoire, et la modération de la plateforme supprime rapidement les contenus à risque, ce qui limite la taille des jeux de données disponibles.

Dans ce contexte de ressources limitées, notre hypothèse principale est que les signaux d'engagement, parce qu'ils sont générés par l'auditoire après consommation de la vidéo, sont moins exposés aux manipulations introduites par le créateur que le texte lui-même. Nous proposons AudiSense, une architecture de fusion multimodale tardive qui combine deux sources d'information complémentaires. La première est un modèle de langage multilingue, XLM-RoBERTa, affiné sur un texte composite regroupant le titre, les mots-clics, la transcription audio et les commentaires. La seconde est un classifieur XGBoost appliqué à quatorze caractéristiques d'engagement, dérivées des cinq compteurs bruts (vues, mentions « j'aime », partages, signets et commentaires) par des ratios d'interaction et un indice de concentration. Les deux probabilités sont combinées par un réseau de portails dynamique qui apprend un poids propre à chaque vidéo, ce qui permet de s'appuyer davantage sur le texte lorsqu'il est clair et sur l'engagement lorsque l'algospeak le dégrade.

Nous avons constitué un corpus de 136 vidéos TikTok et de 1 339 commentaires, que nous avons annotés manuellement avec un accord inter-annotateurs de 0,76. Sur l'ensemble de test, AudiSense atteint une aire sous la courbe précision-rappel de 0,86 et une aire sous la courbe ROC de 0,92, ce qui le place devant les neuf systèmes auxquels nous l'avons comparé, qu'il s'agisse de méthodes lexicales classiques, de modèles séquentiels neuronaux ou de stratégies de fusion plus récentes. L'étude d'ablation est sans doute le résultat le plus parlant : prise seule, la branche d'engagement obtient 0,76 d'AUPRC, alors que la branche textuelle plafonne à 0,48. Autrement dit, quand l'algospeak dégrade le texte, l'engagement reste un signal solide. Le portail dynamique nous a aussi appris quelque chose : il accorde en moyenne un poids de 0,68 au texte pour les vidéos positives et de seulement 0,36 pour les vidéos négatives, ce qui montre qu'il ajuste vraiment sa décision d'une vidéo à l'autre. Comme l'ensemble de test reste petit, nous présentons ces résultats comme ceux d'une étude pilote plutôt que comme des conclusions définitives. À notre connaissance, ce travail propose le premier corpus multimodal annoté et la première fusion adaptative texte-engagement appliqués à la détection d'idéation suicidaire sur TikTok.

Mots-clés : détection d'idéation suicidaire, TikTok, Fusion multimodale, Algospeak, XLM-RoBERTa, XGBoost, Données limitées.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

API	Application Programming Interface
AUC	Area Under the Curve
AUPRC	Area Under the Precision-Recall Curve
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
DGN	Dynamic Gating Network
GPU	Graphics Processing Unit
HHI	Herfindahl-Hirschman Index
LR	Logistic Regression
MCC	Matthews Correlation Coefficient
MLP	Multi-Layer Perceptron
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
TALN	Traitement Automatique du Langage Naturel
TF-IDF	Term Frequency - Inverse Document Frequency
XGBoost	Extreme Gradient Boosting
XLM-RoBERTa	Cross-lingual Language Model RoBERTa

DÉDICACE

À mon père, pour sa confiance, ses encouragements et les valeurs qu'il m'a transmises.

À ma mère, pour son amour, sa patience, ses sacrifices et son soutien indéfectible tout au long de mon parcours.

À mon époux, pour sa présence, sa compréhension, son soutien quotidien et sa patience durant les moments les plus exigeants de cette aventure académique.

À mes frères et sœurs, pour leurs encouragements, leur affection et leur soutien moral.

À ma famille, mes amis et à toutes les personnes qui ont cru en moi et m'ont accompagnée tout au long de ce parcours.

Enfin, à toutes les personnes qui, derrière un écran, traversent en silence des moments difficiles, et à celles qui œuvrent à les accompagner.

REMERCIEMENTS

Ce mémoire est l'aboutissement d'un travail qui n'aurait pas été possible sans le soutien de plusieurs personnes, que je tiens à remercier ici.

Je tiens tout d'abord à remercier mon directeur de recherche, **M. Hamdi Ben Abdessalem**, ainsi que ma codirectrice, **Mme Haïfa Nakouri**, pour leur encadrement, leurs conseils et leur accompagnement tout au long de cette maîtrise. Leurs commentaires et leurs suggestions ont contribué à l'amélioration de ce travail.

Je souhaite également remercier les professeurs du **Département d'informatique et de mathématiques de l'Université du Québec à Chicoutimi**, ainsi que l'ensemble du personnel académique et administratif. Les connaissances acquises au cours de ma formation, la qualité de l'enseignement reçu et l'environnement offert par l'UQAC ont joué un rôle important dans mon parcours universitaire et personnel.

J'adresse mes sincères remerciements à ma famille et à mes amis pour leur soutien, leur confiance et leurs encouragements tout au long de ces années d'études. Je remercie particulièrement **mon époux** pour sa patience, sa compréhension et son soutien constant dans les moments les plus exigeants de ce parcours. Une pensée affectueuse à **ma chère sœur**, dont la présence, les éclats de rire, les blagues et les moments de complicité ont apporté de la légèreté même dans les périodes les plus chargées. Je remercie également **mon frère** qui, malgré la distance qui nous sépare, a toujours su être présent par ses encouragements, son soutien et sa confiance.

Je souhaite enfin adresser un remerciement tout particulier à **ma mère**. Depuis mon enfance, elle a consacré son temps, son énergie et son amour à notre éducation, à mon frère, ma sœur et moi. Par ses sacrifices, sa patience, ses prières et son soutien inconditionnel, elle nous a permis de poursuivre nos rêves et de croire en nos capacités. Aucun mot ne saurait exprimer pleinement ma gratitude pour tout ce qu'elle a fait pour nous. Ce mémoire lui doit beaucoup. Je souhaite également remercier **mon père** pour son attention, son inquiétude bienveillante et sa façon de toujours penser à notre bien-être, qui ont été une source précieuse de réconfort et de motivation tout au long de mon parcours.

Ce travail aborde un sujet sensible. Je tiens à exprimer une pensée particulière pour toutes les personnes confrontées de près ou de loin à l'idéation suicidaire, ainsi qu'aux familles touchées. Si la recherche présentée dans ces pages peut, à terme et de manière même modeste, contribuer à mieux comprendre la manière dont la détresse psychologique s'exprime sur les plateformes numériques, alors l'effort consacré à ces travaux aura atteint sa raison d'être.

AVANT-PROPOS

Ce mémoire présente les travaux que nous avons menés dans le cadre de la maîtrise en informatique à l'Université du Québec à Chicoutimi. Le projet se situe au croisement du traitement automatique du langage naturel et de la santé mentale, avec une attention portée aux plateformes de vidéos courtes, un terrain encore peu exploré par la recherche dans ce domaine.

Le sujet abordé est sensible. La détection d'idéation suicidaire touche à la détresse réelle de personnes, et nous avons gardé cette réalité à l'esprit tout au long du travail. Notre objectif n'est pas de remplacer le jugement humain, mais d'étudier dans quelle mesure des signaux automatiques peuvent aider à repérer des contenus à risque sur une plateforme où ces signaux sont volontairement masqués. Nous restons conscients des limites de ce type d'approche et des précautions qu'impose son éventuel usage.

Nous avons choisi la première personne du pluriel tout au long du document, conformément à l'usage académique francophone. Ce choix reflète aussi le caractère collectif de la recherche, qui a bénéficié de l'encadrement et des relectures de plusieurs personnes.

Si la lecture de ce mémoire, ou une situation personnelle, suscite de la détresse, des ressources d'aide sont disponibles. Au Québec, le service Suicide.ca est accessible en tout temps. À l'international, l'Association internationale pour la prévention du suicide tient un répertoire de centres de crise par pays [1].

1. https://www.iasp.info/resources/Crisis_Centres/

INTRODUCTION GÉNÉRALE

Ce mémoire propose une approche multimodale pour la détection automatique d'idéation suicidaire sur TikTok. Le sujet se situe à l'intersection du traitement automatique du langage naturel, de la santé mentale et de l'analyse des médias sociaux. Si la détection de signaux de détresse psychologique sur les plateformes textuelles classiques telles que Twitter ou Reddit est désormais un domaine de recherche relativement mature, l'extension de ces approches aux plateformes de vidéos courtes soulève des défis spécifiques que la littérature commence seulement à documenter.

Notre travail s'inscrit dans cette extension et s'attaque à un problème encore peu exploré : la détection fiable de l'idéation suicidaire dans un environnement où le texte produit par le créateur de la vidéo est volontairement dégradé par des codes d'obfuscation lexicale, et où les corpus annotés disponibles restent de taille très limitée. Pour répondre à ces contraintes, nous proposons AudiSense, une architecture de fusion multimodale tardive combinant un modèle de langue multilingue affiné pour le texte et un classifieur à arbres boostés pour l'engagement comportemental de l'auditoire, reliés par un mécanisme de portail dynamique apprenant un poids spécifique à chaque vidéo. Cette architecture est évaluée sur un corpus que nous avons constitué et annoté manuellement, qui constitue à notre connaissance le premier ensemble multimodal annoté pour cette tâche sur TikTok.

Le mémoire est organisé en cinq chapitres. Le chapitre 1 précise le contexte, la problématique, les questions de recherche, les hypothèses et les objectifs spécifiques qui guident le travail. Le chapitre 2 examine l'état de l'art sur lequel s'appuie notre démarche. Le chapitre 3 expose la méthodologie, depuis la collecte et l'annotation du corpus jusqu'à l'architecture d'AudiSense et au protocole d'évaluation. Le chapitre 4 présente les résultats expérimentaux et leur discussion. Le chapitre 5 approfondit les limites méthodologiques et les considérations

éthiques, et esquisse les directions de recherche futures. La conclusion générale synthétise les apports du travail et ouvre sur ses prolongements possibles.

CHAPITRE I

PROBLÉMATIQUE ET OBJECTIFS

Dans ce chapitre, nous présentons en détail le contexte qui motive ce travail, les particularités de TikTok comme plateforme de partage de vidéos courtes, le phénomène linguistique de l'algospeak et les contraintes de collecte qui caractérisent ce domaine. Nous formulons ensuite la problématique de recherche, énonçons les hypothèses qui la sous-tendent et précisons les objectifs spécifiques poursuivis. Le chapitre se termine sur les premières considérations éthiques.

1.1 CONTEXTE

1.1.1 LA SANTÉ MENTALE À L'ÈRE DES PLATEFORMES NUMÉRIQUES

L'idéation suicidaire renvoie à l'ensemble des pensées portant sur la possibilité de mettre fin à ses jours, allant du simple souhait de ne plus exister à l'élaboration de plans concrets. Comme le rappellent Ghanadian *et al.* (2025), sa détection à partir des contenus partagés sur les plateformes numériques constitue un domaine actif de recherche en santé publique. Spangenberg *et al.* (2025) soulignent toutefois qu'aucune définition opérationnelle ne fait pleinement consensus dans la littérature clinique, ce qui invite à expliciter, dans tout travail empirique, les critères d'annotation retenus. Sa détection précoce constitue un enjeu majeur de santé publique. Au Canada, l'Agence de la santé publique estime que près de deux cents tentatives sont effectuées chaque jour, dont une fraction conduit à un décès[2]. La pandémie de COVID-19 et les confinements qui l'ont accompagnée ont accentué la visibilité des troubles anxio-dépressifs, sans toujours offrir des canaux d'expression ou de soutien adaptés. Dans

2. https://www.canada.ca/fr/sante-publique/services/prevention-suicide.html

ce contexte, les plateformes numériques sont devenues un lieu où la détresse psychologique s'exprime de manière croissante, parfois ouvertement, parfois de manière implicite à travers des codes communautaires.

Le traitement automatique du langage naturel s'est emparé de cette question depuis une dizaine d'années. Coppersmith *et al.* (2014) ont été parmi les premiers à montrer que les signaux de détresse mentale étaient quantifiables à partir de publications Twitter. Cette ligne de recherche s'est ensuite enrichie de modèles plus sophistiqués, exploitant les représentations contextuelles produites par les architectures fondées sur l'attention. Les travaux successifs de Sawhney *et al.* (2021b), Sawhney *et al.* (2021a) et Sawhney *et al.* (2021c) ont notamment proposé d'intégrer la dynamique temporelle, l'état émotionnel et la topologie sociale dans des classifieurs spécialisés. Plus récemment, des modèles de langage pré-entraînés sur du corpus en santé mentale, tels que MentalBERT (Ji *et al.*, 2022) ou KETCH (Zhang *et al.*, 2025), ont permis d'obtenir des gains supplémentaires sur les corpus de référence.

Néanmoins, l'essentiel de cette recherche s'est concentré sur des plateformes textuelles à formats relativement longs, principalement Twitter et Reddit. Ces plateformes présentent une caractéristique partagée : le texte y est le vecteur principal du message, et les utilisateurs disposent d'un espace suffisant pour développer leurs pensées de manière relativement explicite. L'apparition et la croissance fulgurante des plateformes de vidéos courtes a modifié cette configuration. TikTok, en particulier, présente un ensemble de propriétés qui rendent l'application directe des méthodes existantes problématique.

1.1.2 TIKTOK ET L'EXPRESSION DE LA DÉTRESSE PSYCHOLOGIQUE

TikTok compte aujourd'hui plus d'un milliard d'utilisateurs actifs dans le monde et occupe une place dominante chez les adolescents et les jeunes adultes (Basch *et al.*, 2022). Son

interface, fondée sur un flux algorithmique de vidéos de courte durée, encourage la création et la consommation rapides de contenu. Dans le corpus que nous avons collecté pour ce mémoire, la durée médiane des vidéos est de 22 secondes, avec une distribution s'étalant de 5 secondes à un peu moins de 7 minutes pour les cas extrêmes. Cette brièveté impose des modes d'expression différents de ceux observés sur les plateformes textuelles.

Plusieurs travaux récents ont commencé à documenter la manière dont la détresse psychologique s'exprime sur TikTok. Basch *et al.* (2022) ont conduit une analyse de contenu portant sur des vidéos étiquetées comme relatives à la santé mentale et ont relevé une proportion notable de témoignages personnels, souvent accompagnés de musiques mélancoliques et de filtres visuels caractéristiques. Grant-Allen *et al.* (2025) ont publié une analyse thématique consacrée spécifiquement aux contenus relatifs à l'automutilation et au suicide sur TikTok. Ils montrent que ces publications obéissent à des codes communautaires distincts, dans lesquels la formulation explicite est rare et où l'émotion est portée par des combinaisons multimodales que les pipelines lexicaux standard peinent à interpréter.

Pastro *et al.* (2026), dans une étude récente consacrée à la détection d'idéation suicidaire sur TikTok, soulignent en outre que la modération de la plateforme est particulièrement active sur ces sujets. Les vidéos identifiées par les systèmes automatiques sont supprimées rapidement, parfois en quelques heures, ce qui rend extrêmement difficile la constitution de corpus annotés de grande taille. Les auteurs reconnaissent explicitement le caractère limité de leurs propres jeux de données et invitent la communauté à composer avec cette contrainte de ressources réduites. Cette difficulté n'est pas anecdotique : elle structure l'ensemble des choix méthodologiques disponibles, et c'est sur cette base que nous fondons une partie de nos décisions techniques dans ce mémoire.

1.2 SPÉCIFICITÉS DE LA DÉTECTION SUR TIKTOK

1.2.1 UN TEXTE FRAGMENTÉ ET MULTILINGUE

Sur les plateformes textuelles classiques, le texte associé à une publication est généralement un bloc homogène, rédigé par un seul auteur dans une seule langue. Sur TikTok, la situation est sensiblement différente. Une vidéo est associée à plusieurs sources textuelles, qui ne sont ni produites au même moment, ni par les mêmes personnes, ni nécessairement dans la même langue. Le créateur fournit d'abord un titre court et une chaîne de mots-clics (*hashtags*) destinés à rejoindre certaines communautés ou à contourner la modération. À cela s'ajoute la transcription audio, qui correspond aux paroles prononcées dans la vidéo lorsqu'il y en a ; elle n'est pas fournie nativement par la plateforme et doit être obtenue par un système de reconnaissance automatique de la parole. Enfin, les commentaires de l'auditoire, postés par d'autres utilisateurs après la publication de la vidéo, viennent enrichir ce contenu d'une couche supplémentaire, fréquemment multilingue, dont la tonalité peut différer radicalement de celle du contenu original.

Dans notre corpus, les commentaires associés à une même vidéo se présentent dans plusieurs langues. L'anglais y est majoritaire, mais nous y avons relevé du français, de l'arabe, du chinois ainsi que des langues moins représentées. Cette pluralité linguistique impose le recours à un modèle de langage multilingue, capable de représenter ces contenus dans un espace de représentations commun sans nécessiter de pipelines de traduction préalables (Conneau *et al.*, 2020).

1.2.2 L'ALGOSPEAK COMME DÉFI STRUCTUREL

Le second élément qui distingue TikTok des plateformes textuelles classiques tient à un phénomène linguistique propre aux environnements modérés algorithmiquement, désigné

dans la littérature sous le nom d'algospeak (Steen *et al.*, 2023). Ce terme renvoie à la pratique par laquelle les créateurs remplacent volontairement les mots sensibles par des variantes codées destinées à contourner les filtres de modération automatique. Steen *et al.* (2023) en ont conduit une étude approfondie et montrent que ces substitutions prennent plusieurs formes systématiques.

La première forme est la substitution lexicale : le mot *suicide* est ainsi fréquemment remplacé par *unalive*, et le mot *kill* par *unalive* également ou par des variantes proches. Le terme *sewerslide* sert également de substitut. La seconde forme est la modification typographique : les utilisateurs insèrent des caractères non alphabétiques dans le mot original, comme dans *su!cide* ou *s.u.i.c.i.d.e*, ce qui fragmente l'unité lexicale attendue par les modèles de langage. La troisième forme passe par les mots-clics : des étiquettes en apparence neutres, comme #unalive, #sewerslide ou des compositions plus opaques, deviennent des points de ralliement communautaires identifiables uniquement par les membres avertis. Ces trois mécanismes opèrent en amont du modèle. Ils modifient le texte avant qu'il ne soit traité, et limitent par construction ce que peut récupérer un classifieur, quelle que soit la richesse de son architecture.

Cette observation est cruciale pour notre travail. Si le signal lexical est délibérément dégradé à la source, l'augmentation de la capacité du modèle ne suffit pas à le restaurer. Il devient alors pertinent de chercher d'autres signaux, moins exposés à cette manipulation, et de concevoir un système capable d'arbitrer entre les sources d'information selon leur fiabilité contextuelle.

1.2.3 UNE MODÉRATION QUI REND LA COLLECTE DE DONNÉES ARDUE

Le troisième élément contextuel concerne la disponibilité même des données. La politique de modération de TikTok, conjuguée à la sensibilité des contenus relatifs au suicide,

conduit à la suppression rapide d'une fraction importante des vidéos pertinentes pour notre étude. Grant-Allen *et al.* (2025) et Pastro *et al.* (2026) le soulignent l'un et l'autre : entre le moment où une vidéo est repérée et celui où une équipe de recherche peut la collecter, archiver et annoter, plusieurs heures peuvent suffire à la voir disparaître. Cette dynamique fait du domaine de la détection d'idéation suicidaire sur TikTok un cas typique d'apprentissage à ressources limitées, ce que la littérature anglo-saxonne désigne sous le terme low-resource.

L'apprentissage à ressources limitées constitue un sous-domaine actif du traitement automatique du langage naturel, en particulier pour les applications en santé mentale, où la rareté des annotations expertes est une caractéristique historique (Ji *et al.*, 2022). Plusieurs stratégies ont été explorées pour pallier ce manque : l'apprentissage multitâche, où des tâches auxiliaires servent à enrichir les représentations partagées ; l'utilisation de modèles pré-entraînés sur des corpus génériques massifs, ensuite affinés avec peu de données (Conneau *et al.*, 2020) ; ou encore l'exploitation de signaux non textuels susceptibles de compléter le texte (Liu *et al.*, 2022). Dans notre travail, nous nous inscrivons explicitement dans cette dernière voie, en cherchant à exploiter les signaux comportementaux de l'auditoire comme source complémentaire d'information.

L'aspect low-resource de notre étude doit être pris au sérieux à plusieurs niveaux. Tout d'abord, il limite la taille des ensembles d'entraînement, de validation et de test, ce qui se traduit par des intervalles de confiance plus larges et une marge d'erreur statistique non négligeable. Ensuite, il impose une vigilance particulière sur les choix méthodologiques : tout protocole d'évaluation qui ne préviendrait pas rigoureusement les fuites entre ensembles pourrait fausser l'interprétation des résultats. Enfin, il impose une certaine humilité dans la formulation des conclusions, qui doivent être présentées comme des indications préliminaires plutôt que comme des affirmations définitives.

1.3 PROBLÉMATIQUE DE RECHERCHE

Au croisement des éléments présentés ci-dessus, la détection d'idéation suicidaire sur TikTok soulève deux questions de recherche distinctes mais complémentaires. La première concerne le pouvoir discriminant relatif des modalités disponibles : entre le texte produit par le créateur, qui est exposé à la dégradation par l'algospeak, et l'engagement comportemental généré par l'auditoire, lequel porte le signal le plus utile à la classification ? La seconde concerne l'architecture de combinaison de ces modalités : sachant que la fiabilité relative du texte et de l'engagement peut varier d'une vidéo à l'autre, une fusion adaptative apprenant un poids spécifique à chaque vidéo offre-t-elle un avantage sur une fusion à pondération globale fixe ?

Les approches existantes répondent partiellement à ces enjeux. Pastro *et al.* (2026) se sont appuyés sur la densité de mots-clics comme heuristique de détection, mais cette stratégie reste sensible aux variations introduites par l'algospeak. Nguyen *et al.* (2025) ont conçu un système nommé MTikGuard qui combine images vidéo, transcriptions automatiques et reconnaissance optique de caractères au sein d'une architecture transformer. Leur objectif est toutefois la modération de contenu sensible à destination des mineurs, et non la détection d'idéation suicidaire ; par ailleurs, leur architecture n'intègre pas les signaux d'engagement comportemental. Cui *et al.* (2024) ont récemment proposé une approche fondée sur Whisper et les grands modèles de langage pour l'évaluation du risque suicidaire à partir de la parole, mais ils travaillent dans un contexte clinique et non sur des contenus de plateforme sociale. À notre connaissance, aucun système ne combine simultanément, dans le contexte de TikTok, l'exploitation des signaux comportementaux de l'auditoire comme source primaire de discrimination et un mécanisme d'arbitrage adaptatif entre modalités pour chaque vidéo. C'est ce vide que nous proposons de combler à travers les questions formulées ci-dessous.

Nous formulons donc deux questions de recherche qui guident l'ensemble du travail :

Q1. Entre le texte multilingue produit par le créateur et l'engagement comportemental généré par l'auditoire, quelle modalité porte le pouvoir discriminant le plus important dans un environnement où l'algospeak dégrade le signal lexical ?

Q2. Une fusion adaptative, apprenant un poids spécifique à chaque vidéo, améliore-t-elle la détection d'idéation suicidaire sur TikTok comparativement aux méthodes unimodales et aux stratégies de fusion à pondération globale fixe ?

1.4 HYPOTHÈSES

À chacune des deux questions de recherche correspond une hypothèse opérationnelle. Nous énonçons d'abord chaque hypothèse sous une forme courte et empiriquement testable, puis nous présentons la justification théorique qui la motive.

H1 : Robustesse de l'engagement face à l'algospeak. En présence d'algospeak, les signaux d'engagement de l'auditoire sont plus performants que le texte du créateur pour détecter l'idéation suicidaire sur TikTok.

Cette hypothèse est liée à Q1. Sa justification théorique repose sur une propriété structurelle des signaux d'engagement : les vues, les mentions « j'aime », les partages, les signets et les commentaires sont générés après que les utilisateurs ont consommé et compris la vidéo, indépendamment des artifices lexicaux choisis par le créateur. Là où l'algospeak modifie le texte avant qu'il ne soit traité par un classifieur, il ne modifie pas la réaction comportementale de l'auditoire. Cette propriété est compatible avec la notion de signal manipulation-résistant proposée par Luceri *et al.* (2025). H1 est empiriquement testable par comparaison directe des

performances de la branche textuelle et de la branche d'engagement, prises isolément, sur le même ensemble de test.

H2 : Avantage de la fusion adaptative. Un système de fusion qui apprend un poids spécifique à chaque vidéo dépasse les performances d'un système unimodal et d'une fusion à pondération globale fixe sur la tâche de détection d'idéation suicidaire.

Cette hypothèse est liée à Q2. Sa justification théorique repose sur l'observation que les deux modalités encodent des aspects partiellement distincts du contenu d'une vidéo, et que la fiabilité relative de chacune n'est pas constante d'une vidéo à l'autre. Pour certaines vidéos, le créateur exprime explicitement la détresse dans la transcription audio ou dans les commentaires qu'il a postés lui-même, et le texte y est très informatif. Pour d'autres, le texte est massivement obfusqué par l'algospeak, et l'engagement devient la source d'information dominante. Un mécanisme apprenant un poids adaptatif devrait donc capter cette variabilité, là où une combinaison à poids fixe en serait incapable par construction. H2 est empiriquement testable par comparaison du système proposé avec les baselines unimodales et les fusions statiques.

1.5 OBJECTIFS

1.5.1 OBJECTIF GÉNÉRAL

L'objectif général de ce mémoire est de concevoir, mettre en œuvre et évaluer un système de détection d'idéation suicidaire sur TikTok qui répond aux contraintes identifiées dans la problématique, à savoir la dégradation du signal textuel par l'algospeak, la taille limitée du corpus annoté disponible, et la nécessité d'une évaluation rigoureuse en régime à faible échantillon.

1.5.2 OBJECTIFS SPÉCIFIQUES

De cet objectif général découlent quatre objectifs spécifiques, ordonnés selon leur correspondance avec les questions de recherche et les hypothèses :

1. **Constituer et annoter un corpus.** Constituer et annoter manuellement un corpus de vidéos TikTok couvrant à la fois des contenus à risque d'idéation suicidaire et des contenus de la classe négative. Ce corpus, qui inclut le titre, les mots-clics, les transcriptions audio, les commentaires et les compteurs d'engagement, constitue le support empirique permettant d'aborder Q1 et Q2.
2. **Concevoir une architecture de fusion adaptative.** Concevoir une architecture multimodale combinant un encodeur multilingue affiné pour le texte et un classifieur à arbres boostés pour l'engagement, dont les sorties sont combinées par un mécanisme de portail dynamique apprenant un poids spécifique à chaque vidéo. Cet objectif rend opérationnel le système associé à H2.
3. **Évaluer le pouvoir discriminant des modalités (Q1, H1).** Conduire une étude d'ablation modalité par modalité, afin de quantifier la contribution respective du texte sans transcription, du texte avec transcription et de l'engagement. Cette analyse vise à répondre directement à Q1 en testant empiriquement H1.
4. **Évaluer l'avantage de la fusion adaptative (Q2, H2).** Comparer le système proposé à un ensemble de baselines incluant des méthodes unimodales et des stratégies de fusion à pondération fixe (fusion précoce, empilement linéaire), puis analyser le comportement interne du portail dynamique afin d'éclairer le fonctionnement du mécanisme d'arbitrage. Cette comparaison répond directement à Q2 en testant empiriquement H2.

1.6 CONTRIBUTIONS

À l'issue de ce travail, nous formulons trois contributions principales.

Premièrement, nous présentons un corpus multimodal annoté manuellement pour la détection d'idéation suicidaire sur TikTok, regroupant les sources textuelles propres à la plateforme et les compteurs d'engagement comportemental associés. À notre connaissance, ce corpus constitue à ce jour un des rares ensembles multimodaux de cette nature pour la tâche considérée.

Deuxièmement, nous proposons AudiSense, une architecture de fusion multimodale tardive combinant un encodeur multilingue pour le texte et un classifieur à arbres boostés pour l'engagement, reliés par un mécanisme de portail dynamique apprenant un poids spécifique à chaque vidéo.

Troisièmement, nous conduisons une évaluation comparative qui situe AudiSense par rapport à un ensemble représentatif de méthodes existantes et qui isole, par une étude d'ablation, la contribution de chaque modalité au système.

1.7 CONSIDÉRATIONS ÉTHIQUES PRÉLIMINAIRES

Le sujet abordé dans ce mémoire est sensible et exige une attention particulière. Nous précisons ici les principes qui ont guidé notre travail, et reviendrons sur les considérations éthiques au Chapitre 5, à la lumière des résultats expérimentaux.

Toutes les données utilisées proviennent de vidéos publiquement accessibles sur TikTok, sans interaction directe avec les utilisateurs concernés. Selon l'article 2.2 de l'*Énoncé de politique des trois Conseils : Éthique de la recherche avec des êtres humains* (Conseil de recherches en sciences humaines du Canada *et al.*, 2022), les recherches non intrusives ne

comportant pas d'interaction directe entre le chercheur et d'autres personnes sur Internet, et pour lesquelles il n'y a pas d'attente raisonnable en matière de respect de la vie privée, ne requièrent pas l'évaluation d'un comité d'éthique de la recherche. Nous nous inscrivons strictement dans ce cadre.

Nous reconnaissons néanmoins que la frontière entre données publiques et données sensibles n'est pas toujours claire lorsque celles-ci touchent à la santé mentale. Plusieurs principes ont guidé notre démarche. Tout d'abord, aucun identifiant susceptible de réidentifier un utilisateur n'est conservé dans le corpus utilisé pour l'apprentissage. Les noms d'auteurs et les pseudonymes des commentateurs présents dans le fichier brut ne sont pas utilisés par les modèles et n'apparaissent dans aucun exemple cité dans ce mémoire. Ensuite, les exemples textuels présentés à titre illustratif sont paraphrasés, ce qui constitue une pratique éthique courante dans la recherche en santé mentale sur médias sociaux. Enfin, nous avons accordé une attention particulière à la formulation des résultats et de la discussion, en veillant à ne pas instrumentaliser la détresse d'individus identifiables.

Nous reconnaissons également les risques d'usage détourné d'un système automatisé de détection d'idéation suicidaire. Un système de ce type, déployé sans accompagnement humain, pourrait produire des faux positifs susceptibles de stigmatiser des utilisateurs n'exprimant pas de détresse réelle, ou des faux négatifs susceptibles de manquer des cas authentiques. Dans tous les cas, notre travail ne constitue pas et ne saurait constituer un outil de diagnostic clinique. Il s'agit d'une recherche exploratoire dont les sorties devraient, dans toute application opérationnelle envisageable, être systématiquement intégrées à un dispositif de triage humain, sous la supervision de professionnels qualifiés en santé mentale.

1.8 SYNTHÈSE DU CHAPITRE

Ce chapitre a posé les fondements du travail. Nous avons d'abord situé la détection d'idéation suicidaire dans son contexte de santé publique, puis montré en quoi le passage des plateformes textuelles classiques aux plateformes de vidéos courtes, et notamment à TikTok, modifie substantiellement les hypothèses de travail. La fragmentation et le multilinguisme du texte, le phénomène d'algospeak et la modération agressive de la plateforme s'imposent comme les trois éléments structurants de la problématique. Sur cette base, nous avons formulé deux questions de recherche et deux hypothèses opérationnelles correspondantes, ainsi que quatre objectifs spécifiques alignés sur ces questions, et identifié trois contributions principales. Le chapitre suivant présente la revue de littérature qui prolonge cette analyse en détaillant l'état de l'art sur lequel notre travail s'appuie.

CHAPITRE II
REVUE DE LITTÉRATURE

Dans ce chapitre, nous présentons les travaux antérieurs sur lesquels s'appuie notre recherche. Nous commençons par la détection d'idéation suicidaire sur les plateformes textuelles classiques, puis nous décrivons les modèles de langage pré-entraînés et les algorithmes à arbres boostés que nous utilisons. Nous abordons ensuite les travaux propres à TikTok et aux vidéos courtes, l'apprentissage multimodal et les mécanismes de fusion, ainsi que l'apprentissage en contexte de données limitées. Nous terminons par les métriques adaptées aux jeux de données déséquilibrés et par le positionnement de notre travail.

2.1 DÉTECTION D'IDÉATION SUICIDAIRE SUR LES MÉDIAS SOCIAUX

La détection automatique des signaux de détresse mentale à partir des médias sociaux est un domaine de recherche actif depuis plus d'une décennie. Les premiers travaux de référence se sont appuyés sur Twitter. Coppersmith *et al.* (2014) ont montré qu'il était possible de quantifier des signaux de santé mentale à partir de publications collectées par expressions régulières, en s'appuyant sur des caractéristiques linguistiques simples. Cette approche a ouvert la voie à de nombreuses études exploitant les marqueurs lexicaux, syntaxiques et de fréquence pour identifier la dépression, l'anxiété ou l'idéation suicidaire.

Une partie importante de cette littérature s'est ensuite déplacée vers Reddit, dont la structure en forums thématiques facilite la constitution de corpus étiquetés. Les sous-forums consacrés à la santé mentale fournissent en effet une étiquette implicite, et plusieurs jeux de données de référence en sont issus. Gkotsis *et al.* (2017) ont caractérisé différentes conditions de santé mentale à partir de ce type de contenu en combinant traitement du langage et

apprentissage automatique. La difficulté centrale, soulignée par la plupart de ces travaux, tient à la rareté des annotations fiables : étiqueter correctement un message comme exprimant ou non une idéation suicidaire demande une expertise que peu d'annotateurs possèdent, et la frontière entre détresse, expression figurée et idéation réelle reste souvent ambiguë.

Les approches les plus récentes ont introduit des représentations plus riches. Sawhney *et al.* (2021b) ont proposé une formulation ordinale de la détection d'idéation suicidaire, qui tient compte de la gradation du risque plutôt que de se limiter à une décision binaire. Dans deux travaux complémentaires, les mêmes auteurs ont intégré la dimension émotionnelle à travers des représentations sensibles à la phase émotionnelle (Sawhney *et al.*, 2021a), puis la dimension temporelle et sociale au moyen de représentations d'utilisateurs combinant historique et entourage (Sawhney *et al.*, 2021c). Ces travaux montrent que le contexte, qu'il soit émotionnel, temporel ou social, apporte une information utile au-delà du seul contenu textuel d'un message isolé. Plus récemment encore, Liu *et al.* (2022) ont étudié l'évaluation du risque suicidaire sur Weibo, confirmant que la problématique dépasse le cadre anglophone et concerne des environnements linguistiques variés.

Ces travaux partagent toutefois une hypothèse commune : le texte y est la source d'information principale, et il est supposé exprimé de manière relativement directe. Cette hypothèse devient fragile sur une plateforme où le texte est volontairement masqué, ce qui motive l'exploration de signaux complémentaires dans notre travail.

2.2 MODÈLES DE LANGAGE PRÉ-ENTRAÎNÉS

Les modèles de langage fondés sur l'architecture Transformers ont profondément transformé le traitement automatique du langage naturel. Le modèle BERT Devlin *et al.* (2019), en particulier, a introduit l'idée d'un pré-entraînement bidirectionnel sur de grandes quantités

de texte non annoté, suivi d'un affinage sur une tâche cible avec relativement peu de données. Cette stratégie est particulièrement intéressante dans les domaines à ressources limitées, puisque le modèle possède déjà une certaine compréhension du langage avant même de voir les données de la tâche.

Pour les applications en santé mentale, des variantes spécialisées ont été proposées. Ji *et al.* (2022) ont présenté MentalBERT, un modèle pré-entraîné sur un corpus de publications issues de forums consacrés à la santé mentale. Les auteurs montrent que ce pré-entraînement spécialisé améliore les performances sur plusieurs tâches de détection, dont l'idéation suicidaire, par rapport à un modèle générique. Dans le même esprit, Zhang *et al.* (2025) ont proposé KETCH, qui enrichit les représentations par des connaissances externes pour la détection de contenus liés à la santé mentale. Ces approches confirment l'intérêt d'adapter les modèles au domaine, mais elles restent largement monolingues et centrées sur l'anglais.

Or, le contenu que nous traitons est multilingue par nature. Les commentaires d'une vidéo TikTok peuvent mêler plusieurs langues, et il serait coûteux de construire des pipelines de traduction préalables pour chacune d'elles. Pour cette raison, nous nous tournons vers XLM-RoBERTa, présenté par Conneau *et al.* (2020). Ce modèle est pré-entraîné sur un très grand corpus couvrant une centaine de langues, ce qui lui permet de représenter des textes de langues différentes dans un même espace de représentations. Les auteurs montrent que cet apprentissage interlingue à grande échelle permet d'obtenir de bonnes performances sur des langues variées sans dégradation marquée sur les langues les mieux dotées. Cette propriété est précieuse dans notre contexte, où nous ne pouvons pas présumer de la langue d'un commentaire donné.

2.3 ALGORITHMES À ARBRES DE DÉCISION BOOSTÉS

Si les modèles Transformers dominent le traitement du texte, les méthodes à arbres de décision boostés restent des références solides pour les données tabulaires, c'est-à-dire pour des observations décrites par un ensemble fixe de caractéristiques numériques. Dans notre travail, les signaux d'engagement, une fois transformés en caractéristiques, constituent précisément ce type de données.

L'algorithme XGBoost, introduit par Chen & Guestrin (2016), est l'un des plus utilisés dans ce contexte. Il construit séquentiellement un ensemble d'arbres, chacun corrigeant les erreurs du précédent, et intègre des mécanismes de régularisation qui limitent le surapprentissage. Cette dernière propriété est importante pour nous, car notre jeu de données est petit et le risque de surapprentissage y est élevé. XGBoost offre également une bonne robustesse aux valeurs aberrantes et aux distributions très asymétriques, deux caractéristiques que nous retrouvons dans les compteurs d'engagement, où quelques vidéos virales atteignent des valeurs extrêmes.

L'efficacité de XGBoost dépend toutefois du réglage de ses hyperparamètres. Pour ce réglage, nous nous appuyons sur Optuna, un cadre d'optimisation présenté par Akiba *et al.* (2019). Optuna explore l'espace des hyperparamètres au moyen d'un estimateur de type Parzen arborescent, qui concentre progressivement la recherche sur les régions prometteuses. Cette approche est plus efficace qu'une recherche exhaustive lorsque le nombre d'essais possibles est limité, ce qui correspond à notre situation.

La combinaison de modèles de langage et d'arbres boostés a déjà été explorée dans le domaine de la santé mentale. Mokni & Haoues (2026) ont récemment proposé des modèles hybrides associant BERT et XGBoost pour la prédiction de l'état de santé mentale à partir de texte. Leur travail diffère du nôtre sur un point essentiel : ils appliquent les deux méthodes à

la même modalité textuelle, alors que nous les affectons à deux modalités distinctes, le texte pour le modèle de langage et l'engagement pour les arbres boostés.

2.4 DÉTECTION SUR TIKTOK ET VIDÉOS COURTES

La recherche consacrée spécifiquement à TikTok est récente et encore peu abondante. Basch *et al.* (2022) ont conduit une analyse de contenu transversale portant sur des vidéos relatives à la santé mentale. Leur étude, de nature descriptive, met en évidence la prévalence des témoignages personnels et la fréquence de certains codes visuels et sonores, mais ne propose pas de système de détection automatique.

Grant-Allen *et al.* (2025) ont réalisé une analyse thématique des contenus liés à l'automutilation et au suicide sur TikTok. Ils documentent les codes communautaires propres à ces contenus et soulignent la rareté des formulations explicites, ce qui complique l'application des méthodes lexicales classiques. Leur travail est qualitatif et n'a pas pour objet de construire un classifieur, mais il fournit une description utile du terrain.

Du côté des systèmes automatiques, Pastro *et al.* (2026) ont proposé une approche de détection d'idéation suicidaire sur TikTok exploitant notamment la densité de mots-clics. Ils reconnaissent que la modération agressive de la plateforme limite la taille des jeux de données et que les approches purement lexicales se heurtent à l'obfuscation introduite par l'algospeak. Nguyen *et al.* (2025) ont pour leur part conçu MTikGuard, un système multimodal fondé sur Transformers combinant images vidéo, transcriptions et reconnaissance optique de caractères, dans le but de protéger les mineurs des contenus sensibles. Bien que proche du nôtre par sa nature multimodale, ce système vise une tâche différente et n'exploite pas les signaux d'engagement. Enfin, Shaikh (2025) a étudié la détection de contenus nuisibles multimodaux

à l'aide de BERT, dans un cadre qui partage certaines préoccupations méthodologiques avec le nôtre sans cibler spécifiquement l'idéation suicidaire ni l'engagement comportemental.

Une autre ligne de travaux pertinente concerne l'analyse de la parole. Cui *et al.* (2024) ont proposé une évaluation du risque suicidaire fondée sur Whisper et les grands modèles de langage, dans un contexte clinique. Sun *et al.* (2025) ont par ailleurs présenté un réseau multimodal à fusion dynamique pour la détection du bien-être à partir de la parole. Ces travaux confortent deux choix de notre méthodologie : l'usage de Whisper pour transcrire l'audio des vidéos, et le recours à un mécanisme de fusion dynamique entre modalités.

2.5 APPRENTISSAGE MULTIMODAL ET MÉCANISMES DE FUSION

L'apprentissage multimodal vise à combiner plusieurs sources d'information de natures différentes, par exemple du texte, de l'image et du son. Baltrusaitis *et al.* (2019) en proposent une synthèse et une taxonomie de référence. Ils distinguent notamment plusieurs stratégies de fusion selon le moment où les modalités sont combinées. La fusion précoce consiste à concaténer les représentations des différentes modalités avant la classification finale. Elle est simple, mais elle suppose que les modalités partagent un espace de représentation compatible, ce qui n'est pas garanti lorsque l'une est issue d'un modèle de langage et l'autre de caractéristiques tabulaires. La fusion tardive, à l'inverse, entraîne un classifieur par modalité puis combine leurs sorties. Elle est plus robuste aux différences de nature entre modalités et se prête mieux aux régimes à faible échantillon, car chaque branche peut être optimisée séparément.

Entre ces deux extrêmes, les mécanismes de portail offrent une voie intermédiaire. Plutôt que de fixer une pondération globale entre modalités, un portail apprend à pondérer les contributions en fonction de l'observation courante. Xue & Marculescu (2022) ont proposé DynMM, un réseau multimodal dynamique qui adapte son traitement à chaque entrée. Sun

et al. (2025) ont appliqué une idée analogue à la détection du bien-être à partir de la parole. Ces travaux montrent qu'une pondération adaptative peut améliorer les performances lorsque la fiabilité relative des modalités varie d'une observation à l'autre, ce qui correspond précisément à notre hypothèse H2.

La combinaison de plusieurs classifieurs relève aussi, plus largement, des méthodes d'ensemble. La technique d'empilement, ou stacking, formalisée par Wolpert (1992), consiste à entraîner un méta-modèle sur les sorties de modèles de base. Notre architecture peut se lire comme une variante de cette idée, dans laquelle le méta-modèle est un portail dynamique opérant sur des méta-caractéristiques dérivées des deux experts. Un point d'attention soulevé par la littérature concerne la fuite d'information : si les sorties servant à entraîner le méta-modèle sont produites par des modèles de base ayant déjà vu les mêmes données, le méta-modèle risque de surestimer ses performances. Pour éviter cet écueil, nous construisons les méta-caractéristiques par validation croisée hors-pli, une stratégie que nous détaillons au Chapitre 4.

Enfin, la pertinence des signaux d'engagement comme source d'information complémentaire trouve un appui théorique dans les travaux de Luceri *et al.* (2025), qui proposent de considérer les signatures comportementales d'engagement comme des signaux résistants à la manipulation. Selon cette perspective, les comportements agrégés d'un grand nombre d'utilisateurs sont plus difficiles à falsifier qu'un contenu produit par un acteur unique. Cette idée fonde directement notre hypothèse H1.

2.6 APPRENTISSAGE EN CONTEXTE DE DONNÉES LIMITÉES

La rareté des données annotées est une contrainte structurelle de la recherche en santé mentale, et plus encore dans le cas de TikTok pour les raisons exposées au chapitre précédent. Plusieurs stratégies ont été développées pour composer avec cette contrainte.

La première consiste à transférer la connaissance acquise sur de grands corpus génériques. C'est le principe du pré-entraînement suivi d'un affinage, dont nous avons déjà parlé à propos de BERT et de XLM-RoBERTa 2.2. Le modèle dispose ainsi d'une représentation du langage avant même de voir les données de la tâche, ce qui réduit le volume d'exemples nécessaires.

Une deuxième stratégie est l'apprentissage multitâche, où des tâches auxiliaires liées à la tâche principale servent à enrichir les représentations partagées. Plusieurs travaux ont montré que l'entraînement conjoint de tâches proches en santé mentale, comme la détection de dépression et d'anxiété aux côtés de l'idéation suicidaire, améliore la performance sur la tâche principale, en raison de la proximité entre ces troubles (Ji *et al.*, 2022; Sawhney *et al.*, 2021b). La rareté et le coût des annotations expertes confirment l'intérêt de méthodes économes en données dans ce domaine.

Une troisième stratégie, celle que nous privilégions, consiste à exploiter des signaux non textuels susceptibles de compléter le texte. Lorsque le texte est dégradé ou peu informatif, d'autres sources peuvent porter une part du pouvoir discriminant. Dans notre cas, ces sources sont les compteurs d'engagement, qui sont disponibles pour toutes les vidéos et ne dépendent pas de la qualité du texte. Cette stratégie ne s'oppose pas aux deux précédentes : nous combinons en réalité le transfert (via XLM-RoBERTa) et l'exploitation de signaux non textuels (via l'engagement), au sein d'une architecture conçue pour le régime à faible échantillon.

Travailler en régime de données limitées impose des précautions méthodologiques, en particulier l'usage d'intervalles de confiance pour interpréter les écarts entre systèmes avec prudence.

2.7 MÉTRIQUES POUR DONNÉES DÉSÉQUILIBRÉES

Le choix des métriques d'évaluation est déterminant lorsque les classes sont déséquilibrées, ce qui est le cas ici puisque les vidéos exprimant une idéation suicidaire sont minoritaires. L'exactitude, qui mesure la proportion de prédictions correctes, est trompeuse dans ce cas : un classifieur qui prédirait systématiquement la classe majoritaire obtiendrait une exactitude élevée tout en étant inutile.

Pour cette raison, nous privilégions des métriques mieux adaptées. L'aire sous la courbe de précision-rappel, ou AUPRC, résume le compromis entre précision et rappel sur l'ensemble des seuils de décision. Elle est particulièrement informative lorsque la classe positive est rare, car elle se concentre sur la capacité du modèle à identifier cette classe. L'aire sous la courbe ROC, dont la signification a été formalisée par Hanley & McNeil (1982), complète cette information en mesurant la capacité du modèle à ordonner correctement les exemples positifs et négatifs, indépendamment du seuil.

Nous utilisons également le coefficient de corrélation de Matthews, ou MCC. Chicco & Jurman (2020) montrent que cette métrique offre une évaluation plus équilibrée que le score F1 ou l'exactitude lorsque les classes sont déséquilibrées, car elle prend en compte les quatre catégories de la matrice de confusion. Un score de MCC élevé n'est obtenu que si le modèle réussit bien sur les deux classes à la fois.

Enfin, parce que les sorties des classifieurs ne sont pas nécessairement calibrées sous la distribution de la classe cible, nous ajustons les seuils de décision en tenant compte de

la proportion attendue de positifs. Cette idée d'ajustement des sorties d'un classifieur à de nouvelles probabilités a priori a été formalisée par Saerens *et al.* (2002). Dans notre travail, elle se traduit par une sélection des seuils sur l'ensemble de validation uniquement, afin de ne pas biaiser l'évaluation finale.

2.8 SYNTHÈSE CRITIQUE ET LACUNES IDENTIFIÉES

La revue présentée permet de prendre la mesure de l'état des lieux, mais elle laisse aussi apparaître plusieurs lacunes que notre travail entend combler. Nous les regroupons ici en quatre points, dans l'ordre des chapitres précédents.

D'abord, la quasi-totalité de la littérature sur la détection d'idéation suicidaire repose sur des plateformes textuelles à formats relativement longs, principalement Twitter et Reddit. Les méthodes qui y ont fait leurs preuves, qu'il s'agisse de représentations contextuelles, d'apprentissage multitâche ou de modèles spécialisés en santé mentale, supposent toutes que le texte est la source primaire d'information et qu'il est exprimé de manière relativement directe. Cette hypothèse devient fragile sur une plateforme où le texte est dégradé en amont par l'algospeak, et où il est par ailleurs fragmenté entre plusieurs canaux multilingues. Aucun des travaux que nous avons recensés ne traite simultanément ces deux propriétés.

Ensuite, la recherche propre à TikTok est encore peu abondante et reste essentiellement descriptive. Les analyses qualitatives existantes documentent les codes communautaires et les pratiques d'algospeak, mais elles ne proposent pas de systèmes automatiques capables d'en tenir compte. Les rares systèmes automatiques publiés pour TikTok exploitent soit la densité de mots-clics comme heuristique, soit la fusion d'images vidéo et de transcriptions pour des tâches de modération distinctes de la détection d'idéation suicidaire. Aucun n'exploite les signaux d'engagement comportemental en tant que source d'information primaire.

Troisièmement, la littérature sur la fusion multimodale offre des outils intéressants, notamment les mécanismes de portail dynamique, mais ces outils sont presque toujours déployés sur des modalités hétérogènes de type texte plus image ou texte plus parole, où chaque branche est elle-même un modèle de haute capacité. Leur application à un contexte où l'une des branches est un classifieur tabulaire opérant sur des compteurs d'engagement n'a, à notre connaissance, pas été très explorée. Cette combinaison est pourtant cohérente avec la nature même des données disponibles sur TikTok et avec l'hypothèse de robustesse de l'engagement face à l'algospeak.

Enfin, le sous-domaine de l'apprentissage en ressources limitées en santé mentale propose des stratégies, comme le transfert depuis de grands corpus génériques ou l'apprentissage multitâche, mais il sous-exploite la possibilité d'introduire une modalité complémentaire et structurellement différente du texte. Lorsque le texte est dégradé à la source, augmenter la quantité d'entraînement textuel ne suffit pas à restaurer le signal perdu. Il devient alors pertinent de chercher l'information ailleurs, dans des signaux qui ne dépendent pas de la production du créateur.

Ces quatre lacunes convergent vers une même conclusion. Il existe un espace de recherche non couvert au croisement de la détection d'idéation suicidaire sur vidéos courtes, de la fusion multimodale adaptative, et de l'apprentissage en ressources limitées. C'est dans cet espace que se situe notre contribution, dont la section suivante précise les contours.

2.9 POSITIONNEMENT DE NOTRE TRAVAIL

La revue qui précède permet de situer précisément notre contribution. La détection d'idéation suicidaire sur les médias sociaux est un domaine établi, mais largement centré sur des plateformes textuelles et sur l'anglais. Les modèles de langage multilingues, comme

XLM-RoBERTa, offrent une réponse au caractère plurilingue de TikTok. Les arbres boostés, comme XGBoost, conviennent au traitement des signaux d'engagement tabulaires. Les travaux propres à TikTok restent peu nombreux et, lorsqu'ils proposent un système automatique, ils n'exploitent pas conjointement le texte multilingue, l'engagement comportemental et un mécanisme d'arbitrage adaptatif. Les méthodes de fusion multimodale et de portail dynamique fournissent le cadre architectural dont nous avons besoin, et la littérature sur les données limitées en justifie les précautions méthodologiques.

Notre travail se place à l'intersection de ces différentes lignes de recherche. Il combine un modèle de langage multilingue et un classifieur à arbres boostés à travers un portail dynamique, dans le but explicite de répondre à la dégradation du signal textuel par l'algospeak et à la contrainte de données limitées propre à TikTok. Le chapitre suivant détaille la méthodologie qui met en œuvre ce positionnement.

CHAPITRE III
MÉTHODOLOGIE

Dans ce chapitre, nous présentons la méthodologie suivie pour construire et évaluer le système proposé. Nous énonçons d'abord les hypothèses à vérifier, puis nous décrivons la collecte des données, leur annotation et leur prétraitement. Nous détaillons ensuite l'ingénierie des caractéristiques d'engagement, la stratégie de découpage des données, et l'architecture d'AudiSense. Nous terminons par la présentation des systèmes de référence et du protocole expérimental. L'ensemble du pipeline est résumé à la Figure 3.1, qui en présente une vue d'ensemble en six étapes. Les données sont d'abord collectées sur TikTok puis annotées manuellement. Elles passent ensuite par une chaîne de prétraitement qui produit, d'une part, un texte composite multilingue et, d'autre part, un vecteur de quatorze caractéristiques d'engagement. Ces deux entrées sont fournies à deux encodeurs distincts, XLM-RoBERTa pour le texte et XGBoost pour l'engagement, dont les probabilités sont assemblées en un vecteur de méta-caractéristiques. Ce vecteur sert d'entrée à un réseau de portail dynamique, qui produit un coefficient α par vidéo et calcule la prédiction finale comme combinaison convexe des deux probabilités. Le seuil de décision, calibré sur l'ensemble de validation, transforme cette probabilité en décision binaire.

3.1 HYPOTHÈSES À VÉRIFIER

La démarche expérimentale vise à vérifier les deux hypothèses formulées au Chapitre 1. Nous les rappelons brièvement, en les reliant aux expériences qui permettront de les évaluer.

La première hypothèse, H1, postule qu'en présence d'algospeak, les signaux d'engagement de l'auditoire sont plus performants que le texte du créateur pour détecter l'idéation

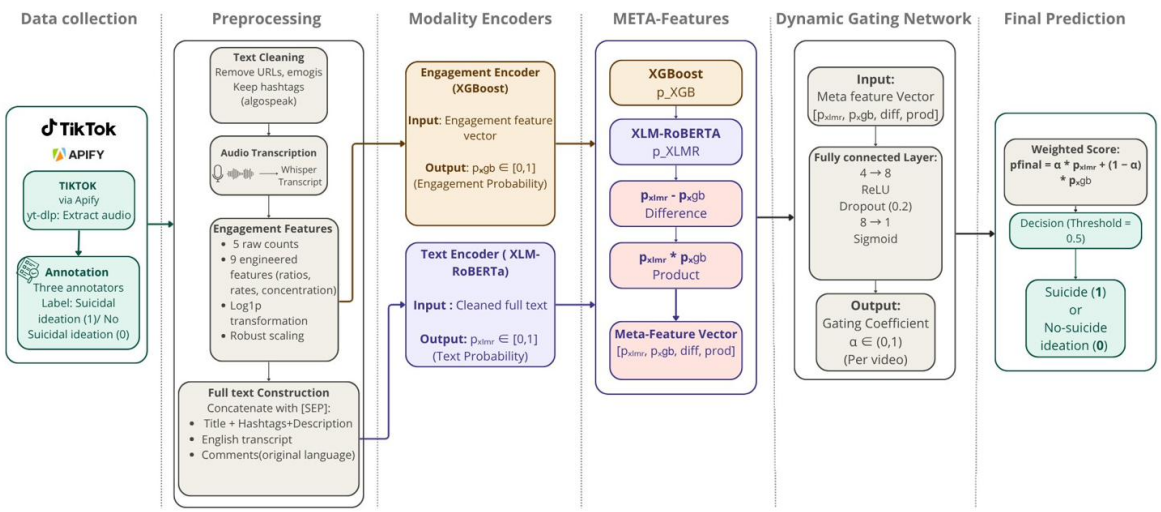

FIGURE 3.1 : Vue d'ensemble du pipeline d'AudiSense.

suicidaire. Pour l'évaluer, nous comparons directement les performances de la branche d'engagement et de la branche textuelle, prises isolément, à travers l'étude d'ablation modalité par modalité.

La seconde hypothèse, H2, postule qu'un système de fusion apprenant un poids spécifique à chaque vidéo dépasse les performances des systèmes unimodaux et des fusions à pondération globale fixe. Elle est évaluée par la comparaison entre AudiSense et l'ensemble des baselines (unimodales, fusion précoce par perceptron, empilement linéaire), complétée par l'analyse de la distribution des poids α appris par le portail, qui éclaire le comportement adaptatif du mécanisme d'arbitrage.

3.2 COLLECTE DES DONNÉES

3.2.1 SOURCE ET OUTILS

Les données proviennent de vidéos publiquement accessibles sur TikTok. La collecte a été réalisée de manière continue depuis l'hiver 2024-2025 jusqu'au début du mois de février 2026. Les vidéos retenues couvrent une plage de publication étendue, allant d'avril 2021 à février 2026, ce qui assure une certaine diversité temporelle dans les styles de contenu et les codes communautaires.

La récupération des vidéos et de leurs métadonnées s'appuie sur la plateforme Apify, qui permet d'extraire de façon automatisée les informations associées à une publication. Pour chaque vidéo, nous récupérons le titre, la date de publication, la chaîne de mots-clics, les cinq compteurs d'engagement (vues, mentions « j'aime », partages, signets et commentaires) ainsi qu'un sous-ensemble de commentaires de l'auditoire. La piste audio de chaque vidéo est extraite à l'aide de l'utilitaire yt-dlp[3], puis convertie au format requis par l'étape de transcription.

3.2.2 TRANSCRIPTION AUDIO

Le contenu parlé des vidéos constitue une source d'information potentiellement utile, mais il n'est pas fourni nativement par la plateforme. Nous obtenons une transcription textuelle à l'aide du système de reconnaissance automatique de la parole Whisper (Radford *et al.*, 2023). Whisper présente deux propriétés intéressantes pour notre contexte. D'une part, il est robuste à la diversité des conditions d'enregistrement, fréquente dans les vidéos amateur. D'autre part, il propose un mode de traduction qui produit directement une transcription en anglais à partir

3. Code disponible sur https://github.com/Ath99/AudiSense/tree/main

d'un audio dans une autre langue. Nous utilisons ce mode afin d'obtenir une transcription homogène en anglais, ce qui facilite son intégration ultérieure au texte composite.

Il convient de noter que la transcription automatique n'est pas exempte d'erreurs, particulièrement lorsque la qualité sonore est faible, lorsque la musique couvre la voix, ou lorsque la vidéo ne contient pas de parole intelligible. Nous reviendrons sur l'effet de ces limites dans la discussion.

3.2.3 DESCRIPTION DU CORPUS

Après agrégation au niveau vidéo, le corpus final comprend 136 vidéos et 1 339 commentaires associés. Chaque vidéo est accompagnée de son titre, de ses mots-clics, de la transcription de sa piste audio et d'un ensemble de commentaires dont le nombre varie de 3 à 20, avec une médiane de 10 commentaires par vidéo. La durée des vidéos est généralement courte, avec une médiane de 22 secondes, ce qui est conforme au format dominant de la plateforme.

Les commentaires se présentent dans plusieurs langues, l'anglais étant majoritaire, suivi du français, de l'arabe et du chinois, parmi d'autres. Cette diversité linguistique confirme la pertinence d'un encodeur multilingue. Les mots-clics, quant à eux, mêlent des étiquettes génériques de visibilité, des étiquettes thématiques liées à la santé mentale et, dans certains cas, des variantes relevant de l'algospeak.

3.3 ANNOTATION DES DONNÉES

3.3.1 SCHÉMA D'ANNOTATION

La tâche est formulée comme une classification binaire au niveau de la vidéo. Une vidéo reçoit l'étiquette 1 lorsqu'elle présente des indices d'idéation suicidaire, et l'étiquette 0 dans le cas contraire. Dans le prolongement des définitions utilisées par les travaux récents en détection automatique (Ghanadian *et al.*, 2025), nous considérons qu'une vidéo relève de la classe positive lorsqu'elle exprime, directement ou par l'intermédiaire de son contenu et des réactions qu'elle suscite, la planification ou le souhait de mettre fin à ses jours. Les contenus relevant de la simple tristesse, de l'expression figurée ou du témoignage rétrospectif sans détresse actuelle sont rangés dans la classe négative, sauf indices contraires.

L'étiquette d'une vidéo est déterminée à partir de l'ensemble de ses éléments. Dans le traitement, lorsque plusieurs lignes de commentaires sont associées à une même vidéo, l'étiquette retenue au niveau vidéo correspond au maximum des étiquettes : il suffit qu'un élément indique un risque pour que la vidéo soit considérée comme positive. Ce choix reflète une logique de prudence, où l'on préfère ne pas manquer un cas potentiel.

3.3.2 ACCORD INTER-ANNOTATEURS

L'annotation a été réalisée par trois personnes, à savoir le premier auteur et les deux personnes assurant la supervision du projet. Chaque vidéo a été examinée indépendamment, puis les désaccords ont été discutés afin d'aboutir à une étiquette consensuelle. Pour quantifier la fiabilité de l'annotation, nous avons calculé le coefficient kappa de Cohen, qui mesure l'accord entre annotateurs en corrigeant l'accord attendu par le hasard. Nous obtenons une valeur de 0,76, ce qui correspond à un accord substantiel selon l'échelle proposée par Landis & Koch (1977). Cette valeur indique une cohérence raisonnable du schéma d'annotation, tout

en laissant transparaître la difficulté intrinsèque de la tâche, déjà relevée dans la littérature sur la détection d'idéation suicidaire.

3.3.3 DISTRIBUTION DES CLASSES

La répartition des étiquettes au niveau vidéo est présentée à la Figure 3.2. Sur les 136 vidéos, 42 sont étiquetées comme positives, soit une proportion d'environ 31 %, et 94 sont étiquetées comme négatives. Cette distribution déséquilibrée est représentative des conditions réelles, où les contenus à risque demeurent minoritaires. Elle justifie le recours à des métriques adaptées au déséquilibre, ainsi qu'à des stratégies de pondération des classes lors de l'entraînement.

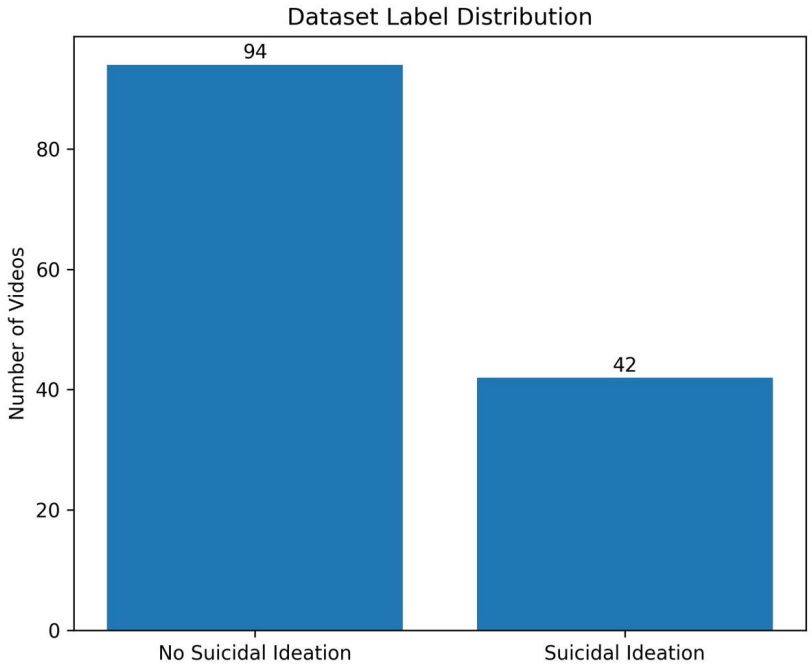

FIGURE 3.2 : Distribution des étiquettes au niveau vidéo dans le corpus.

3.4 PRÉTRAITEMENT DU TEXTE

Le texte associé à chaque vidéo provient de quatre sources : le titre, les mots-clics, la transcription audio et les commentaires de l'auditoire. Ces sources sont d'abord nettoyées, puis assemblées en une entrée textuelle unique.

3.4.1 NETTOYAGE DE DONNÉES

Le nettoyage du texte suit une procédure volontairement légère, afin de ne pas détruire l'information liée à l'algospeak. Concrètement, le texte est mis en minuscules, les adresses URL sont supprimées, et la ponctuation est retirée à l'exception du caractère dièse. Ce dernier point est important : les mots-clics constituent un vecteur privilégié de l'algospeak, et conserver le marqueur dièse permet au modèle de les distinguer du texte courant. Les espaces multiples sont enfin normalisés. À la différence des chaînes de prétraitement classiques utilisées pour les méthodes lexicales, nous n'appliquons ni racinisation, ni lemmatisation, ni suppression de mots vides, car le modèle de langage pré-entraîné est conçu pour traiter directement le texte brut.

3.4.2 CONSTRUCTION DU TEXTE COMPOSITE

Une fois nettoyées, les sources textuelles sont concaténées en une entrée unique. L'ordre adopté place d'abord le titre et les mots-clics, puis la transcription audio, puis les commentaires de l'auditoire, ces blocs étant séparés par un marqueur spécial. Les commentaires associés à une même vidéo sont eux-mêmes joints par ce marqueur. Cette construction permet au modèle de traiter conjointement le contenu produit par le créateur et les réactions de l'auditoire au sein d'une même séquence.

Pour les besoins de l'étude d'ablation, nous construisons également une variante du texte composite excluant la transcription audio. La comparaison entre les deux variantes permet d'isoler la contribution propre du transcript.

3.5 INGÉNIERIE DES CARACTÉRISTIQUES D'ENGAGEMENT

L'engagement d'une vidéo est décrit à l'origine par cinq compteurs bruts : le nombre de vues, de mentions « j'aime », de partages, de signets et de commentaires. Ces compteurs présentent deux difficultés. D'une part, ils sont fortement corrélés à la popularité globale d'une vidéo plutôt qu'à son contenu : une vidéo virale accumule beaucoup d'interactions de toute nature, indépendamment de son sujet. D'autre part, leur distribution est extrêmement asymétrique, quelques vidéos atteignant des valeurs des milliers de fois supérieures à la médiane.

3.5.1 EXTRACTION DE CARACTÉRISTIQUES

Pour atténuer ces deux problèmes : d'une part la corrélation avec la popularité de la vidéo plutôt qu'avec son contenu, et d'autre part l'asymétrie extrême de la distribution, nous enrichissons les cinq compteurs bruts de neuf caractéristiques dérivées, ce qui porte à quatorze le nombre total de caractéristiques d'engagement. Les caractéristiques dérivées se répartissent en trois groupes.

Le premier groupe rassemble six ratios d'interaction, qui normalisent un compteur par un autre afin de capturer un comportement relatif plutôt qu'un volume absolu. Il s'agit du ratio de mentions « j'aime » par vue, de commentaires par vue, de partages par vue, de signets par vue, de commentaires par mention « j'aime » et de signets par partage. Ces ratios reflètent l'intensité de la réaction de l'auditoire indépendamment de la portée de la vidéo.

Le deuxième groupe comprend deux mesures agrégées : l'engagement total, défini comme la somme des mentions « j'aime », des commentaires, des partages et des signets, et le taux d'engagement, défini comme cet engagement total rapporté au nombre de vues.

Le troisième groupe est constitué d'un indice de concentration de l'engagement, adapté de l'indice de Herfindahl-Hirschman (Nasiri *et al.*, 2022). Pour une vidéo donnée, cet indice est calculé comme suit :

$$\mathrm{HHI} = \sum_{i=1}^{4} s_i^2 \tag{3.1}$$

où s_i représente la part du canal i (mentions « j'aime », commentaires, partages, signets) dans l'engagement total de la vidéo. Un HHI proche de 1 indique que l'engagement est concentré sur un seul canal, tandis qu'une valeur proche de 0,25 indique une répartition homogène entre les quatre canaux. L'ensemble des quatorze caractéristiques d'engagement, organisé en trois groupes, est récapitulé dans le Tableau 3.1.

TABLEAU 3.1 : Les quatorze caractéristiques d'engagement.

Groupe	Caractéristique	Description
Compteurs bruts (5)	View_count	Nombre de vues
	Like_count	Mentions « j'aime »
	Comment_count	Commentaires
	Share_count	Partages
	Bookmark_count	Signets
Ratios d'interaction (6)	Like_per_view	Mentions « j'aime » par vue
	comment_per_view	Commentaires par vue
	share_per_view	Partages par vue
	bookmark_per_view	Signets par vue
	comment_per_like	Commentaires par mention « j'aime »
	bookmark_per_share	Signets par partage
Agrégats et concentration (3)	eng_total	Somme des quatre canaux d'engagement
	eng_rate	Engagement total par vue
	eng_hhi	Indice de Herfindahl-Hirschman

3.5.2 TRANSFORMATIONS

Les caractéristiques numériques à forte asymétrie font l'objet d'une transformation logarithmique de type `log1p`, qui comprime les valeurs extrêmes tout en préservant l'ordre. La Figure 3.3 illustre l'effet de cette transformation sur la distribution de plusieurs caractéristiques : les distributions initialement très étalées deviennent plus régulières, ce qui facilite l'apprentissage.

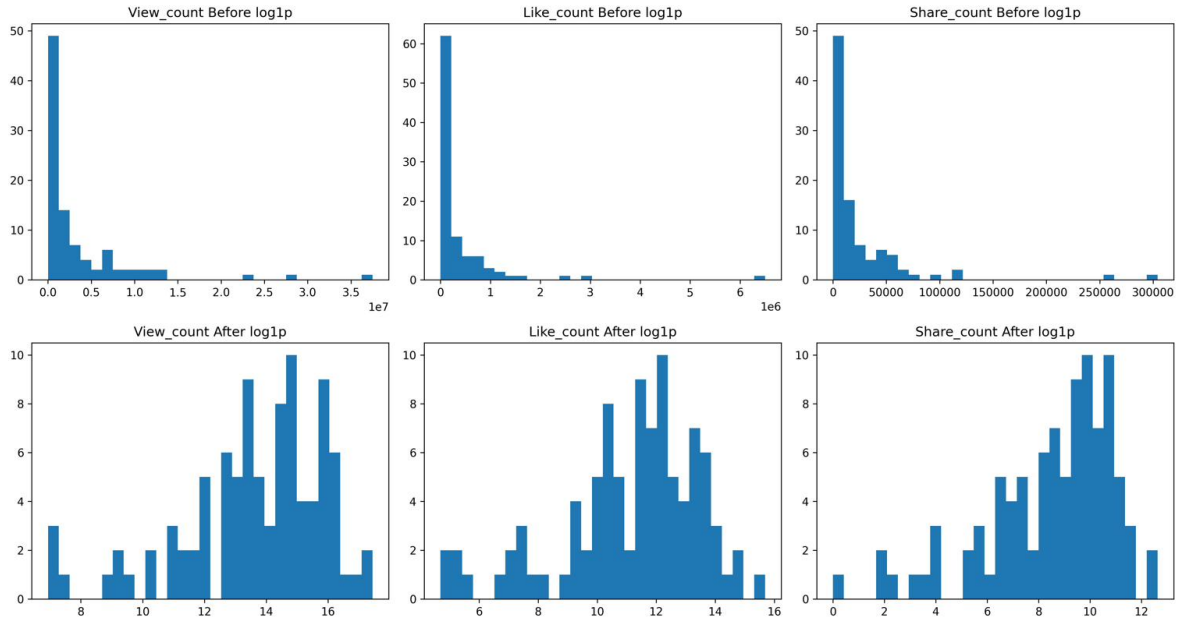

FIGURE 3.3 : Effet de la transformation logarithmique sur la distribution des caractéristiques d'engagement.

Les caractéristiques sont ensuite mises à l'échelle au moyen d'un *RobustScaler*, qui centre et réduit les valeurs en s'appuyant sur la médiane et l'écart interquartile plutôt que sur la moyenne et l'écart-type. Ce choix renforce la robustesse aux valeurs aberrantes résiduelles. Un point méthodologique mérite d'être souligné : la mise à l'échelle est ajustée uniquement sur l'ensemble d'entraînement, puis appliquée telle quelle aux ensembles de validation et de test.

Les valeurs manquantes éventuelles sont par ailleurs imputées par la médiane de l'ensemble d'entraînement, calculée elle aussi avant la mise à l'échelle. Ces précautions évitent toute fuite d'information depuis les ensembles de validation et de test vers l'apprentissage.

La Figure 3.4 présente la matrice de corrélation entre les caractéristiques d'engagement. Elle révèle que certaines caractéristiques sont fortement corrélées entre elles, ce qui est attendu compte tenu de leur mode de construction. Les arbres boostés que nous utilisons tolèrent toutefois bien cette redondance, contrairement à des modèles linéaires qui en seraient davantage affectés.

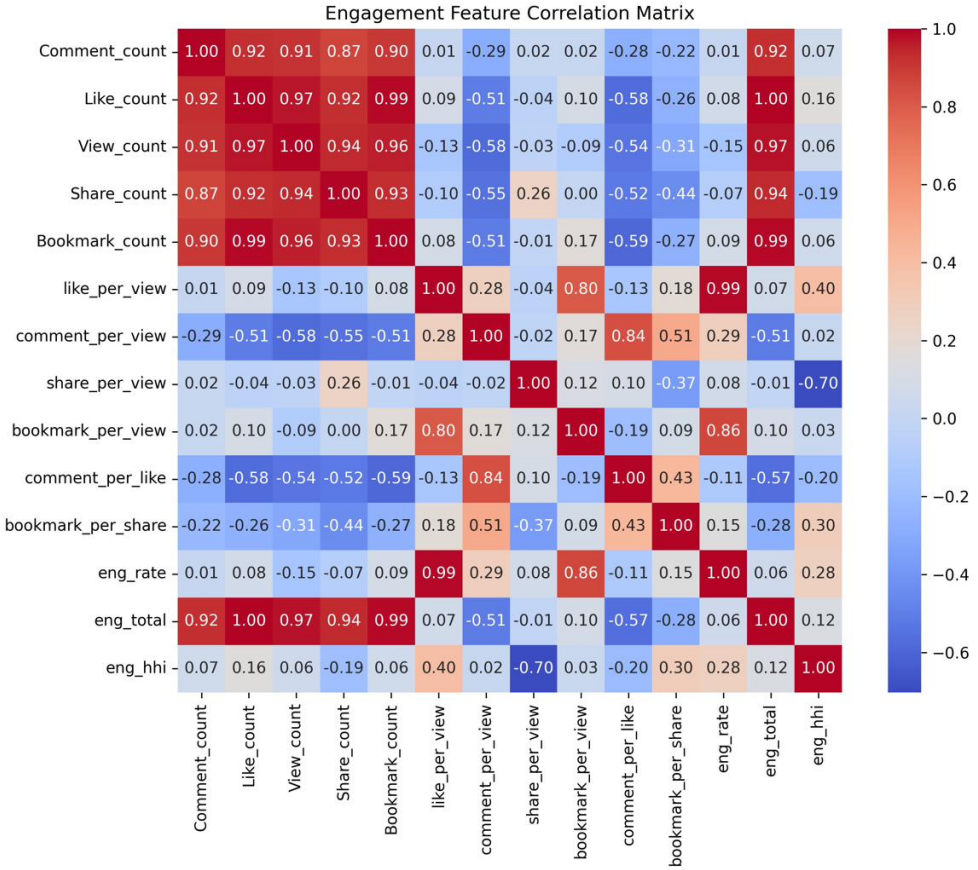

FIGURE 3.4 : Matrice de corrélation entre les caractéristiques d'engagement.

3.6 STRATÉGIE DE VALIDATION

Le découpage des données en ensembles d'entraînement, de validation et de test obéit à deux principes. Le premier est la stratification : les proportions de chaque classe sont préservées dans chaque ensemble, ce qui est essentiel lorsque la classe positive est minoritaire. Le second est le découpage au niveau vidéo : toutes les lignes associées à une même vidéo sont placées dans le même ensemble, afin d'éviter qu'un commentaire d'une vidéo se retrouve dans l'apprentissage tandis qu'un autre commentaire de la même vidéo servirait au test. Sans cette précaution, une fuite d'information surviendrait et fausserait l'évaluation.

Le découpage suit un ratio approximatif de 70 % pour l'entraînement, 15 % pour la validation et 15 % pour le test. Concrètement, 70 % des vidéos sont d'abord affectées à l'entraînement, puis les 30 % restantes sont partagées à parts égales entre validation et test. Ce partage aboutit à 95 vidéos d'entraînement, 20 de validation et 21 de test. L'ensemble de test contient sept vidéos positives. La Figure 3.5 illustre cette répartition. Le découpage est figé par une graine aléatoire afin d'assurer la reproductibilité des expériences.

3.7 ARCHITECTURE PROPOSÉE : AUDISENSE

AudiSense repose sur une fusion multimodale tardive de deux experts entraînés séparément, dont les sorties sont combinées par un réseau de portail dynamique. Nous décrivons successivement chacune des composantes.

3.7.1 EXPERT TEXTUEL

L'expert textuel est un modèle XLM-RoBERTa, dans sa version de base, affiné sur la tâche de classification binaire. Ce modèle prend en entrée le texte composite décrit précé-

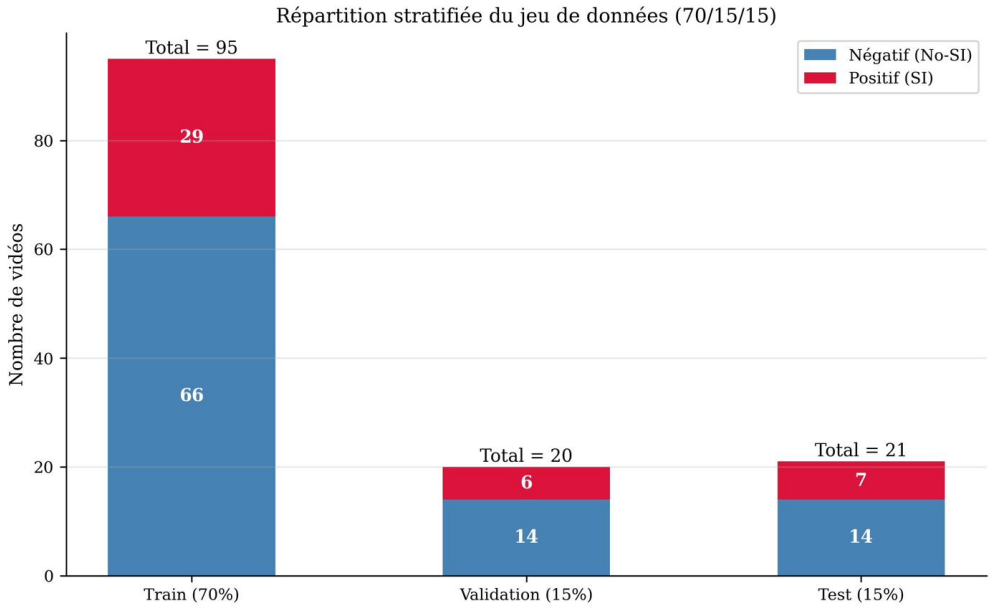

FIGURE 3.5 : Répartition stratifiée des vidéos entre les ensembles d'entraînement, de validation et de test.

demment, tronqué à une longueur maximale de 512 unités lexicales. Il produit en sortie une probabilité d'appartenance à la classe positive, que nous notons p_{texte}. Le caractère multilingue du modèle lui permet de représenter dans un même espace les commentaires de langues différentes, sans étape de traduction préalable.

Pour tenir compte du déséquilibre des classes, la fonction de perte d'entraînement est pondérée en faveur de la classe positive. Le poids appliqué à cette classe n'est pas fixé arbitrairement : il est sélectionné parmi un ensemble de valeurs candidates, et la valeur retenue est celle qui maximise la performance estimée par validation croisée à cinq plis sur l'ensemble d'entraînement uniquement. Ce choix évite d'utiliser l'ensemble de validation pour ce réglage, ce qui préserverait son rôle d'évaluation indépendante.

3.7.2 EXPERT D'ENGAGEMENT

L'expert d'engagement est un classifieur XGBoost (Chen & Guestrin, 2016) appliqué aux quatorze caractéristiques d'engagement. Il produit lui aussi une probabilité d'appartenance à la classe positive, notée p_{eng}. Ses hyperparamètres sont réglés à l'aide d'Optuna (Akiba *et al.*, 2019), par une étude de quarante essais utilisant une validation croisée à cinq plis sur le seul ensemble d'entraînement. La métrique optimisée lors de ce réglage est l'aire sous la courbe de précision-rappel, cohérente avec le déséquilibre des classes. Comme pour l'expert textuel, le réglage n'utilise jamais l'ensemble de validation, afin d'éviter toute fuite.

3.7.3 CONSTRUCTION DES MÉTA-CARACTÉRISTIQUES

Les deux experts produisent chacun une probabilité. Plutôt que de fournir directement ces deux valeurs au portail, nous construisons un vecteur de méta-caractéristiques à quatre dimensions, composé des deux probabilités, de leur différence et de leur produit. La différence renseigne sur le désaccord entre les deux experts, tandis que le produit indique leur accord conjoint. Ce vecteur fournit au portail une information plus riche que les seules probabilités brutes, tout en restant de faible dimension, ce qui convient au régime à faible échantillon.

La manière dont ces méta-caractéristiques sont calculées diffère selon l'ensemble considéré, afin d'éviter une fuite d'information. Pour l'ensemble d'entraînement, les probabilités des deux experts sont obtenues par prédiction hors-pli : chaque expert est entraîné sur une partie des données et prédit sur la partie laissée de côté, de sorte qu'aucun expert ne prédit sur des exemples qu'il a vus pendant son entraînement. Pour les ensembles de validation et de test, les probabilités sont produites par les modèles finaux entraînés sur l'ensemble complet d'entraînement. Cette distinction respecte le protocole d'empilement correct décrit par Wolpert (1992).

3.7.4 RÉSEAU DE PORTAIL DYNAMIQUE

Le cœur d'AudiSense est le réseau de portail dynamique. Il prend en entrée le vecteur de méta-caractéristiques à quatre dimensions et produit un coefficient de pondération α compris entre 0 et 1, propre à chaque vidéo. La prédiction finale est alors une combinaison convexe des deux probabilités d'experts comme le montre l'Équation 3.2 :

$$\hat{p} = \alpha \cdot p_{\text{texte}} + (1 - \alpha) \cdot p_{\text{eng}}. \tag{3.2}$$

Lorsque α est proche de 1, la prédiction s'appuie principalement sur l'expert textuel ; lorsque α est proche de 0, elle s'appuie principalement sur l'expert d'engagement. Le portail apprend ainsi à doser la contribution de chaque modalité selon le contexte de la vidéo, ce qui constitue la traduction concrète de l'hypothèse H2.

Le coefficient α est produit par un petit perceptron multicouche, composé d'une première couche linéaire qui projette les quatre méta-caractéristiques vers huit dimensions, suivie d'une activation de type *ReLU*, d'un *dropout* de probabilité 0,2 servant à limiter le surapprentissage, puis d'une seconde couche linéaire ramenant à une seule dimension, et enfin d'une fonction sigmoïde qui contraint la sortie dans l'intervalle ouvert entre 0 et 1.

L'entraînement du portail utilise une fonction de perte focale (Lin *et al.*, 2017), qui concentre l'apprentissage sur les exemples difficiles et atténue l'effet du déséquilibre des classes. Nous fixons ses deux paramètres aux valeurs usuelles, soit un facteur de pondération de 0,8 et un facteur de focalisation de 2,0. L'optimisation est conduite avec l'algorithme AdamW Loshchilov & Hutter (2017), assorti d'un taux d'apprentissage de 0,01 et d'une régularisation de poids de l'ordre de 10^{-3}, sur 150 itérations.

3.7.5 TAILLE DU PORTAIL

Le réseau de portail est volontairement petit : environ 100 paramètres, contre des centaines de milliers pour XLM-RoBERTa. Ce choix s'explique par la contrainte de données, puisque le portail est entraîné sur seulement 115 exemples (95 d'entraînement et 20 de validation). Un réseau plus large risquerait de mémoriser ces exemples au lieu d'apprendre une règle généralisable.

À cette échelle, on peut se demander si un portail neuronal apporte vraiment quelque chose par rapport à un empilement linéaire plus simple. Pour le vérifier, nous avons inclus un empilement par régression logistique (Wolpert, 1992) parmi les systèmes de référence, avec les mêmes méta-caractéristiques en entrée. Les résultats du chapitre 4 montrent qu'il reste en retrait du portail, ce qui justifie ce choix d'architecture malgré sa petite taille.

3.8 APPROCHES DE RÉFÉRENCE

Afin de situer la performance d'AudiSense, nous le comparons à un ensemble d'approches de référence couvrant plusieurs familles de méthodes. Toutes sont évaluées sous le même protocole, avec calibration des seuils sur l'ensemble de validation ou sur les prédictions hors-pli, jamais sur l'ensemble de test.

La première famille rassemble des méthodes lexicales classiques, fondées sur une représentation TF-IDF (Salton & McGill, 1983) du texte suivie d'une régression logistique (Cox, 1958) ou d'une machine à vecteurs de support (Cortes & Vapnik, 1995). La deuxième famille comprend un modèle séquentiel neuronal de type BiLSTM (Hochreiter & Schmidhuber, 1997), initialisé avec les plongements GloVe-300 (Pennington *et al.*, 2014), qui traite le texte comme une suite ordonnée de mots. La troisième famille regroupe les experts unimodaux pris isolément : l'expert textuel XLM-RoBERTa seul, et l'expert d'engagement XGBoost seul. La

quatrième famille rassemble des stratégies de fusion alternatives : une fusion précoce par perceptron multicouche, qui concatène les représentations des deux modalités, et un empilement linéaire par régression logistique sur les mêmes méta-caractéristiques que celles utilisées par AudiSense. Nous incluons enfin une variante hybride associant un encodeur de type BERT et XGBoost, inspirée des travaux récents de Mokni & Haoues (2026).

Cette diversité de systèmes de référence permet d'isoler l'apport de chacune des composantes d'AudiSense : l'apport du modèle de langage par rapport aux méthodes lexicales, l'apport de la multimodalité par rapport aux experts isolés, et l'apport du portail dynamique par rapport aux fusions statiques.

3.9 PROTOCOLE EXPÉRIMENTAL

3.9.1 MÉTRIQUES D'ÉVALUATION

Conformément à la discussion du chapitre précédent, la métrique principale est l'aire sous la courbe de précision-rappel, mieux adaptée que l'exactitude au déséquilibre des classes. Nous rapportons également l'aire sous la courbe ROC (Hanley & McNeil, 1982), le score F1 de la classe positive, sa précision, son rappel, ainsi que le coefficient de corrélation de Matthews (Chicco & Jurman, 2020). Ce dernier offre une mesure équilibrée tenant compte des quatre cases de la matrice de confusion.

3.9.2 CALIBRATION DU SEUIL DE DÉCISION

Les modèles produisent des probabilités, qu'il faut convertir en décisions binaires à l'aide d'un seuil. Plutôt que de fixer ce seuil à la valeur usuelle de 0,5, nous le choisissons de façon à maximiser le score F1 sur un ensemble réservé, à savoir l'ensemble de validation ou les prédictions hors-pli, selon le système considéré. Le seuil n'est jamais ajusté sur l'ensemble

de test, ce qui garantit que l'évaluation finale reste indépendante. Cette procédure s'appuie sur le principe d'ajustement des sorties d'un classifieur à la distribution cible (Saerens *et al.*, 2002).

3.9.3 INTERVALLES DE CONFIANCE

Compte tenu de la petite taille de l'ensemble de test, une mesure ponctuelle de performance serait trompeuse. Nous accompagnons donc chaque métrique d'un intervalle de confiance à 95 %, obtenu par rééchantillonnage bootstrap. Concrètement, nous tirons mille échantillons avec remise à partir de l'ensemble de test, recalculons la métrique sur chacun, et rapportons les bornes correspondant aux centiles 2,5 et 97,5 de la distribution obtenue. Cette procédure fournit une estimation honnête de l'incertitude associée à chaque mesure, et invite à interpréter les écarts entre systèmes avec la prudence requise.

3.10 SYNTHÈSE DU CHAPITRE

Ce chapitre a décrit l'ensemble de la démarche méthodologique, depuis la collecte des vidéos jusqu'au protocole d'évaluation. Nous avons présenté la constitution et l'annotation du corpus, le prétraitement du texte préservant les indices d'algospeak, l'ingénierie des quatorze caractéristiques d'engagement, et la stratégie de découpage stratifié au niveau vidéo. Nous avons ensuite détaillé l'architecture d'AudiSense, articulée autour de deux experts et d'un portail dynamique opérant sur des méta-caractéristiques calibrées, en soulignant les précautions prises contre les fuites d'information et la prudence imposée par le faible échantillon. Le chapitre suivant présente les résultats expérimentaux obtenus à l'issue de ce protocole et en propose une discussion.

CHAPITRE IV
RÉSULTATS ET DISCUSSION

Dans ce chapitre, nous présentons les résultats expérimentaux obtenus à l'issue du protocole décrit aux chapitres précédents. Nous commençons par les performances d'AudiSense sur l'ensemble de test, puis nous le comparons à ses deux experts unimodaux et à sept systèmes de référence. Nous présentons ensuite l'étude d'ablation, l'analyse de l'importance des caractéristiques d'engagement, le comportement du portail dynamique et l'analyse des erreurs. Nous terminons par une synthèse comparative et par une discussion sur la portée des résultats.

Les résultats qui suivent sont présentés comme une étude pilote, pour les raisons discutées au Chapitre 5. À cette échelle, aucune différence ponctuelle entre deux systèmes ne peut atteindre la significativité statistique au sens strict. Les résultats présentés ci-dessous doivent donc être lus comme des éléments préliminaires issus d'une étude pilote.

4.1 PERFORMANCE D'AUDISENSE

Sur l'ensemble de test, AudiSense obtient une aire sous la courbe de précision-rappel de 0,86, une aire sous la courbe ROC de 0,92, un score F1 de la classe positive de 0,75 (pour une précision de 0,67 et un rappel de 0,86) et un coefficient de corrélation de Matthews de 0,61. Le seuil de décision, calibré au préalable sur les prédictions hors-pli du pool d'entraînement et de validation, est de 0,49. Pour rendre compte de l'incertitude associée à la petite taille de l'ensemble de test, nous avons également calculé des intervalles de confiance par rééchantillonnage bootstrap à 95 %. À titre indicatif, celui de l'AUPRC d'AudiSense s'étend de 0,56 à 1,00, et celui du coefficient de Matthews de 0,24 à 0,91. Ces intervalles, larges en raison du

faible effectif, confirment que les chiffres ponctuels rapportés ici doivent être interprétés avec prudence.

Le rappel relativement élevé indique que le système identifie correctement la majorité des vidéos positives, ce qui est précieux dans un contexte de détection d'idéation suicidaire, où les faux négatifs sont particulièrement coûteux. La précision plus modérée signifie en revanche qu'environ une vidéo prédite positive sur trois est en réalité négative, un compromis cohérent avec la calibration du seuil sur la métrique F1.

4.2 STATISTIQUES FINALES DES DONNÉES

Avant de présenter les résultats expérimentaux, nous récapitulons ici les caractéristiques quantitatives du corpus utilisé et de son partitionnement.

4.2.1 CORPUS COMPLET

Le corpus final regroupe 136 vidéos TikTok et 1 339 commentaires associés, collectés entre avril 2021 et février 2026. Chaque vidéo est décrite par un ensemble de variables textuelles (titre, mots-clics, transcription audio, commentaires de l'auditoire) et un ensemble de quatorze caractéristiques d'engagement (Tableau 3.1). L'annotation manuelle de l'intention communicative a abouti à 42 vidéos positives, soit 30,9 % du corpus, et 94 vidéos négatives, soit 69,1 %. L'accord inter-annotateurs sur l'échantillon doublement annoté atteint un coefficient $\kappa = 0{,}76$.

4.2.2 DÉCOUPAGE DES DONNÉES

Le partitionnement est effectué au niveau de la vidéo afin d'éviter toute fuite entre ensembles, selon un ratio approximatif 70/15/15 stratifié sur l'étiquette de classe. Le Tableau 4.1 récapitule les tailles obtenues.

TABLEAU 4.1 : Répartition des 136 vidéos entre les ensembles d'entraînement, de validation et de test.

Ensemble	Vidéos positives	Vidéos négatives	Total
Entraînement	29	66	95
Validation	6	14	20
Test	7	14	21
Total	**42**	**94**	**136**

Une graine aléatoire (`SEED = 42`) garantit la reproductibilité du partitionnement. La proportion de la classe positive est préservée dans chaque ensemble : 30,5 % à l'entraînement, 30,0 % à la validation et 33,3 % au test.

4.3 HYPERPARAMÈTRES OPTIMAUX

Les deux experts unimodaux d'AudiSense, ainsi que les baselines hybrides, font l'objet d'une recherche d'hyperparamètres par Optuna, conduite sur l'ensemble d'entraînement par validation croisée stratifiée. Le critère d'optimisation est l'AUPRC, qui privilégie le rappel de la classe positive minoritaire. Le Tableau 4.2 rassemble les valeurs retenues pour les principaux composants du système.

La fonction de perte du portail est une perte focale (*focal loss*) paramétrée par $\alpha = 0{,}8$ et $\gamma = 2{,}0$, valeurs choisies pour compenser le déséquilibre de classes (les vidéos positives sont minoritaires) et augmenter le poids des exemples que le modèle classe mal en cours

TABLEAU 4.2 : Hyperparamètres optimaux retenus après optimisation par Optuna.

Composant	Hyperparamètre	Valeur retenue
XLM-RoBERTa (expert textuel)	Taux d'apprentissage	2×10^{-5}
	Taille de lot	8
	Nombre d'époques	5
	Longueur de séquence	256 jetons
XGBoost (expert d'engagement)	`n_estimators`	300
	`max_depth`	4
	`learning_rate`	0,016
	`subsample`	0,69
	`colsample_bytree`	0,84
	`min_child_weight`	7
	`gamma`	2,64
	`scale_pos_weight`	3,11
Portail dynamique (DGN)	Optimiseur	AdamW
	Taux d'apprentissage	0,01
	Régularisation	10^{-3}
	Nombre d'itérations	150

d'entraînement. Le seuil de décision final, calibré sur l'ensemble de validation, est fixé à 0,4922.

4.4 COMPARAISON AVEC LES SYSTÈMES DE RÉFÉRENCE

Le Tableau 4.3 présente les performances d'AudiSense aux côtés de ses deux experts unimodaux et des sept systèmes de référence présentés à la Section 3.8 : la régression logistique (Cox, 1958) et la machine à vecteurs de support (Cortes & Vapnik, 1995) sur TF-IDF (Salton & McGill, 1983), le BiLSTM (Hochreiter & Schmidhuber, 1997) avec plongements GloVe (Pennington *et al.*, 2014), la fusion précoce par perceptron multicouche, l'empilement logistique (Wolpert, 1992), la fusion de caractéristiques de Liu *et al.* (2022) et le système hybride BERT + XGBoost de Mokni & Haoues (2026). Les seuils de décision ont été calibrés pour chaque système sur l'ensemble de validation ou sur les prédictions hors-pli, jamais sur l'ensemble de test.

TABLEAU 4.3 : Comparaison sur l'ensemble de test.

Système	AUPRC	ROC-AUC	F1 (pos)	Précision	Rappel	MCC
AudiSense (approche proposée)	**0,86**	**0,92**	**0,75**	**0,67**	**0,86**	**0,61**
XLM-RoBERTa (texte)	0,48	0,68	0,59	0,50	0,71	0,34
XGBoost (engagement)	0,76	0,88	0,36	0,50	0,29	0,17
Empilement logistique	0,76	0,88	0,36	0,50	0,29	0,17
Fusion précoce (MLP)	0,29	0,36	0,43	0,31	0,71	-0,08
TF-IDF + LR	0,48	0,56	0,45	0,33	0,71	0,00
TF-IDF + SVM	0,50	0,58	0,45	0,33	0,71	0,00
BiLSTM (GloVe-300)	0,46	0,64	0,20	0,33	0,14	0,00
BERT + XGB hybride	0,33	0,46	0,35	0,30	0,43	-0,07
Fusion de caractéristiques	0,48	0,64	0,22	0,50	0,14	0,11

Ce tableau montre qu'AudiSense obtient les meilleures valeurs ponctuelles sur les quatre métriques considérées. L'écart entre AudiSense et le deuxième système, l'empilement logistique, est de 0,09 point d'AUPRC, mais cet écart se traduit surtout dans le score F1 (0,75 contre 0,36) et le coefficient de Matthews (0,61 contre 0,17), deux métriques plus sensibles à l'équilibre entre précision et rappel. Cela suggère que le portail dynamique apporte un bénéfice particulier sur le compromis entre les deux types d'erreur.

Deuxième observation, l'expert d'engagement seul, XGBoost, atteint déjà un AUPRC de 0,76, soit nettement au-dessus de l'expert textuel, qui se situe à 0,48. Ce constat est cohérent avec l'hypothèse H1 selon laquelle l'engagement, généré par l'auditoire après consommation de la vidéo, serait moins exposé aux dégradations de l'algospeak que le texte produit par le créateur.

Troisième observation, les méthodes lexicales classiques restent modestes. La régression logistique et la machine à vecteurs de support sur TF-IDF obtiennent des AUPRC de l'ordre de 0,48 à 0,50. Le BiLSTM avec plongements GloVe se situe au même niveau, autour de 0,46. La proximité de ces trois valeurs avec celle de l'expert XLM-RoBERTa textuel suggère que,

sur ce corpus, le plafond imposé par l'algospeak est commun à toutes les approches purement textuelles.

Quatrième observation, plusieurs systèmes de fusion alternatifs obtiennent des performances très inférieures à celles d'AudiSense. La fusion précoce par MLP se révèle particulièrement défaillante, avec un AUPRC de 0,29 et un coefficient de Matthews légèrement négatif. Cette chute s'explique sans doute par la disparité des échelles entre la probabilité textuelle, comprise dans l'intervalle unitaire, et les quatorze caractéristiques d'engagement transformées : concaténer brutalement ces deux espaces semble nuisible. Le système hybride BERT + XGB inspiré de Mokni & Haoues (2026) obtient également des performances faibles, ce qui peut tenir au fait que ses plongements de dimension élevée sont mal ajustés au régime à faible échantillon.

Enfin, l'empilement logistique obtient des chiffres rigoureusement identiques à ceux de l'expert d'engagement seul, ce qui mérite d'être commenté. Sur cet ensemble de test, il semble que la régression logistique du méta-modèle ait appris à se reposer presque exclusivement sur la probabilité de la branche d'engagement. Ce phénomène illustre concrètement la difficulté d'un méta-modèle linéaire à arbitrer entre deux sources dont les distributions diffèrent, et il motive a posteriori le choix d'un portail dynamique capable de moduler la pondération par vidéo.

4.5 COURBES ROC ET PRÉCISION-RAPPEL

Les courbes ROC et précision-rappel des cinq principaux systèmes sont présentées aux Figures 4.1 et 4.2. Elles donnent une vue continue de la performance, indépendante du seuil de décision particulier retenu.

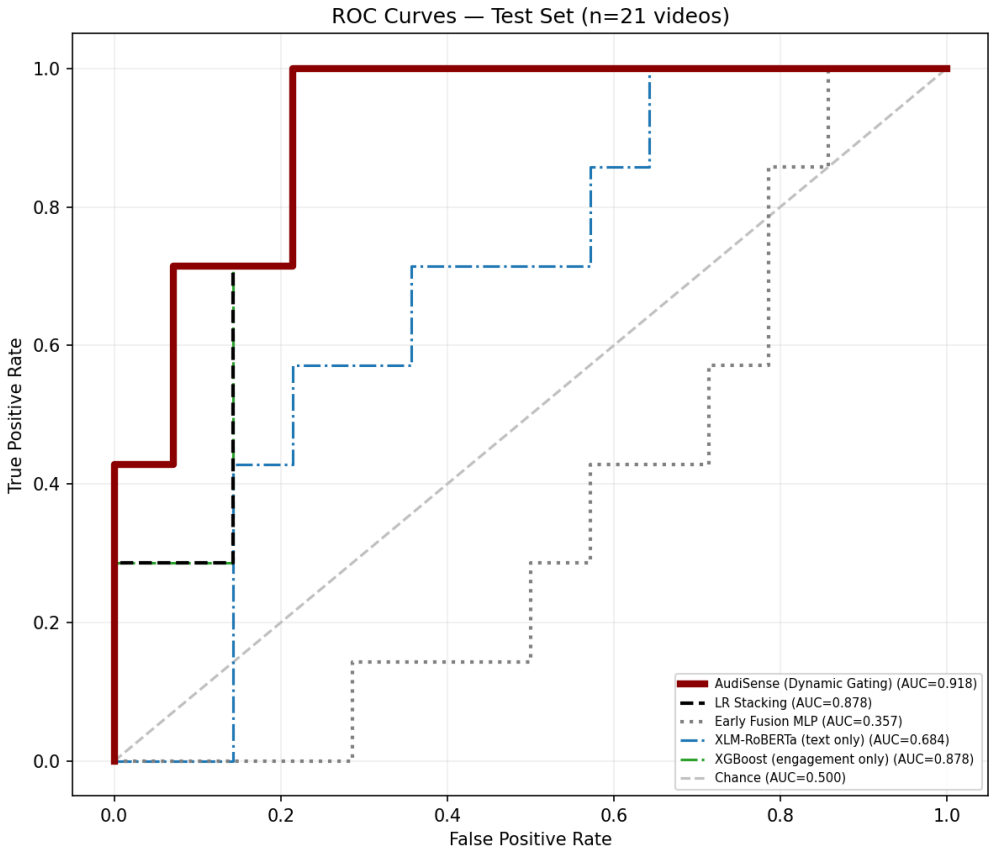

FIGURE 4.1 : Courbes ROC sur l'ensemble de test.

La courbe ROC d'AudiSense domine l'espace lorsqu'on parcourt les seuils. Elle atteint un taux de vrais positifs d'environ 0,70 pour un taux de faux positifs inférieur à 0,10, ce qu'aucun autre système ne parvient à faire. La courbe ROC de l'expert d'engagement et celle de l'empilement logistique sont superposées, ce qui confirme la remarque faite à la section précédente. L'expert textuel XLM-RoBERTa se situe nettement en deçà, et la fusion précoce par MLP descend même sous la ligne de hasard sur une partie de la plage de seuils, ce qui en fait un système contre-productif sur cet ensemble de test.

La courbe précision-rappel donne une lecture complémentaire. Elle est mieux adaptée au déséquilibre des classes, car elle ne récompense pas les vrais négatifs, qui dominent ici.

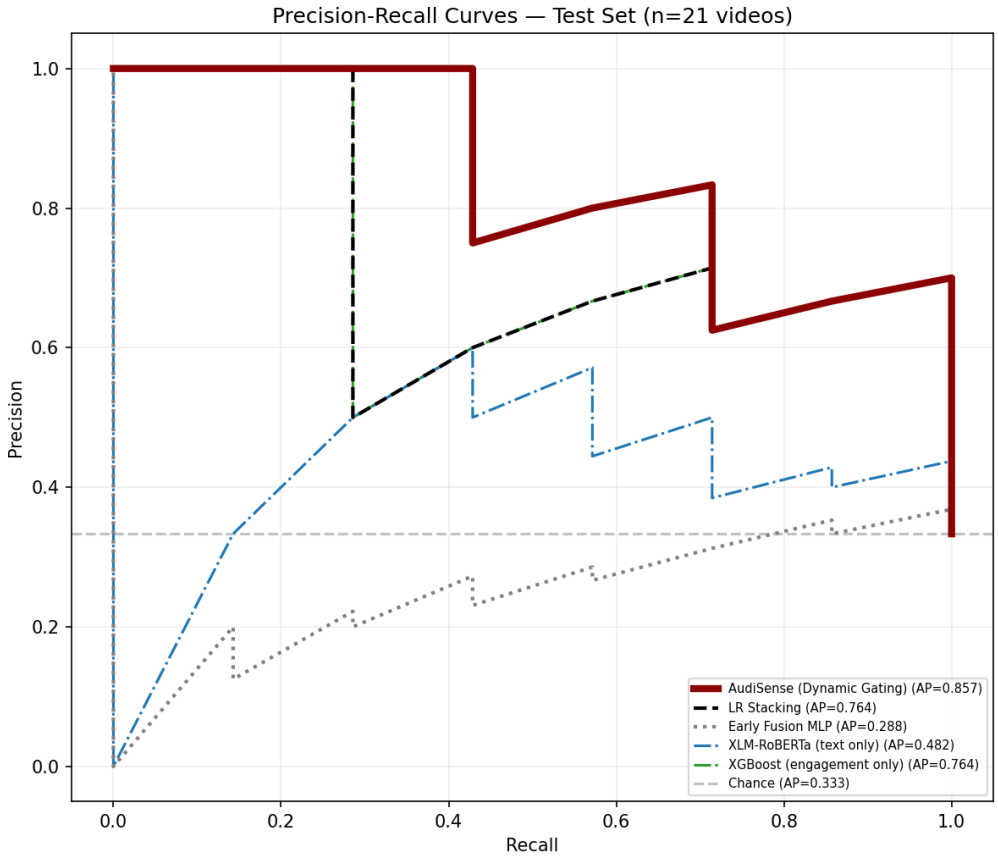

FIGURE 4.2 : Courbes précision-rappel sur l'ensemble de test.

La courbe d'AudiSense reste au-dessus des autres sur une large plage de rappel et conserve une précision supérieure à 0,60 jusqu'à un rappel d'environ 0,70. L'expert d'engagement et l'empilement logistique sont à nouveau superposés, juste en dessous d'AudiSense. La fusion précoce, ici encore, reste très basse, à peine au-dessus de la ligne de hasard donnée par la proportion de positifs sur l'ensemble de test.

4.6 ÉTUDE D'ABLATION

L'étude d'ablation, présentée dans le Tableau 4.4, isole la contribution respective des différentes sources d'information. Quatre conditions sont comparées : AudiSense complet,

l'expert d'engagement seul, l'expert textuel avec transcription audio et l'expert textuel sans transcription audio.

TABLEAU 4.4 : Étude d'ablation sur l'ensemble de test. Comparaison des contributions par modalité.

Condition	AUPRC	ROC-AUC	F1 (pos)	Précision	Rappel
Approche proposée (texte + transcript + engagement)	0,86	0,92	0,75	0,67	0,86
Engagement seul (XGBoost)	0,76	0,88	0,36	0,50	0,29
Texte complet (titre + tags + transcript + commentaires)	0,48	0,68	0,59	0,50	0,71
Texte sans transcription audio	0,45	0,56	0,50	—	—

Trois observations principales ressortent de cette ablation.

Premièrement, l'écart entre la branche d'engagement seule, à 0,76 d'AUPRC, et la branche textuelle complète, à 0,48, est substantiel. La branche d'engagement est manifestement le contributeur dominant dans notre architecture, et ce constat soutient l'hypothèse H1. Les signaux comportementaux générés par l'auditoire portent dans ce corpus une information discriminante plus robuste que celle qu'on peut extraire du texte produit par le créateur.

Deuxièmement, le passage de la branche textuelle sans transcription à la branche textuelle avec transcription apporte un gain modeste, de 0,45 à 0,48 d'AUPRC, soit trois centièmes. La transcription audio par Whisper apporte donc une information utile, mais relativement marginale à l'échelle de notre corpus. Plusieurs facteurs peuvent expliquer cette modestie : la qualité variable des transcriptions, la présence fréquente de musique de fond qui couvre la voix, et l'absence de parole intelligible dans certaines vidéos.

Troisièmement, la fusion des deux modalités, qui porte AudiSense de 0,76 à 0,86 d'AUPRC, indique un gain par rapport à la meilleure modalité prise isolément. Ce gain de dix points illustre la complémentarité des deux sources d'information : le texte, même imparfait, apporte une information qui se combine utilement à l'engagement, à condition que

la combinaison soit adaptée au contexte de chaque vidéo, ce que permet le portail dynamique. Cette observation soutient ainsi l'hypothèse H2 sur l'avantage de la fusion adaptative.

4.7 IMPORTANCE DES CARACTÉRISTIQUES D'ENGAGEMENT

L'importance des quatorze caractéristiques d'engagement, telle qu'estimée par XGBoost à partir du gain moyen apporté par chaque variable dans les arbres, est représentée à la Figure 4.3. Cette analyse clarifie la nature du signal capté par la branche d'engagement.

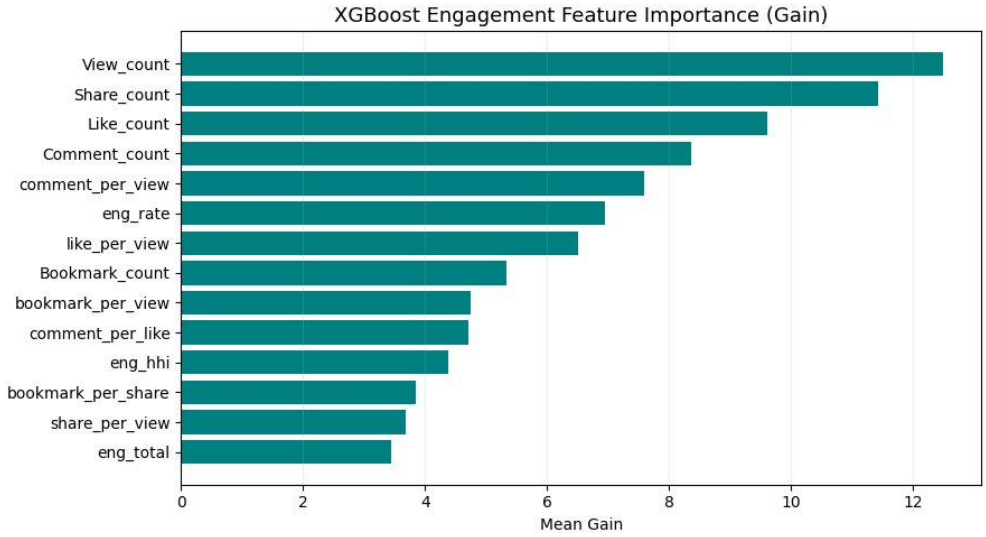

FIGURE 4.3 : Importance des caractéristiques d'engagement selon le gain XGBoost.

Les compteurs bruts dominent le classement selon le gain XGBoost. En tête figurent le nombre de vues (View_count) et le nombre de partages (Share_count), suivis du nombre de mentions « j'aime » (Like_count) et du nombre de commentaires (Comment_count). Ce résultat peut sembler surprenant au regard de leur dépendance à la popularité globale d'une vidéo, mais il reflète probablement leur forte variance dans le corpus, qui les rend mécaniquement informatifs pour la séparation des classes. Parmi les caractéristiques dérivées, le ratio de commentaires par vue (comment_per_view) et le taux d'engagement global (eng_-

rate) affichent les gains les plus élevés, capturant l'intensité relative de la réaction de l'auditoire indépendamment de la portée de la vidéo. L'indice de concentration (eng_hhi) contribue de façon plus modeste, en bas du classement avec bookmark_per_share, share_per_view et eng_total.

Ces résultats ne remettent pas en cause l'intérêt des caractéristiques dérivées, qui restent nécessaires pour normaliser l'effet de popularité et permettre une comparaison équitable entre vidéos de portées très différentes. Les vidéos exprimant une idéation suicidaire ne se distinguent pas nécessairement par un volume absolu d'interactions différent, mais leur profil d'engagement présente des particularités que les ratios et indices dérivés permettent de capturer, comme un ratio élevé de commentaires par mention « j'aime », pouvant traduire un engagement réflexif et empathique de l'auditoire qualitativement différent du divertissement passif.

4.8 COMPORTEMENT DU PORTAIL DYNAMIQUE

Une question importante reste à examiner : le portail dynamique se contente-t-il d'apprendre une pondération globale, ou ajuste-t-il réellement son comportement à chaque vidéo ? La Figure 4.4 présente la distribution du coefficient α produit par le portail sur les vingt-et-une vidéos de l'ensemble de test, en fonction de leur vraie étiquette.

Le coefficient α moyen sur l'ensemble de test est de 0,46. La répartition est toutefois bimodale et clairement liée à l'étiquette. Pour les vidéos positives, le coefficient moyen est de 0,68, indiquant que le système accorde plus de poids à la branche textuelle. Pour les vidéos négatives, le coefficient moyen tombe à 0,36, indiquant que le système s'appuie davantage sur la branche d'engagement.

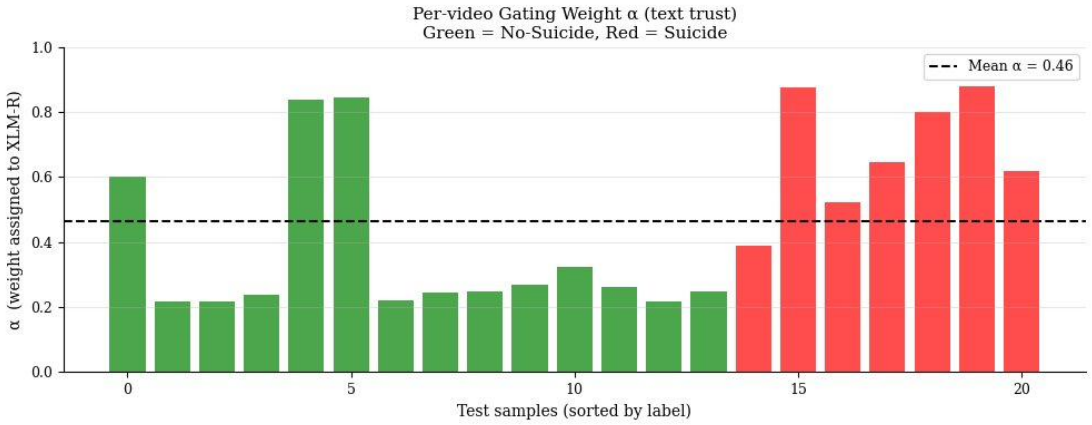

FIGURE 4.4 : Distribution du coefficient de portail α par vidéo sur l'ensemble de test.

Cette différence est compatible avec l'hypothèse H2. Lorsqu'une vidéo contient un texte explicitement chargé en marqueurs de détresse, par exemple lorsque le créateur ou ses commentateurs s'expriment sans recourir à l'algospeak, le portail apprend à faire confiance au texte. À l'inverse, lorsque le texte est dilué ou bénin, le portail se rabat sur les signaux d'engagement, plus stables. Le système ne se réduit donc pas à une combinaison à pondération fixe : il module sa décision selon le contenu observé.

Pour les raisons exposées au début de ce chapitre (taille limitée de l'ensemble de test, intervalles de confiance larges), ce comportement reste à confirmer sur un corpus de taille substantiellement supérieure.

4.9 ANALYSE DES ERREURS

La matrice de confusion d'AudiSense sur l'ensemble de test est présentée à la Figure 4.5. Elle décompose les vingt-et-une prédictions en quatre cases : vrais positifs, faux positifs, vrais négatifs et faux négatifs.

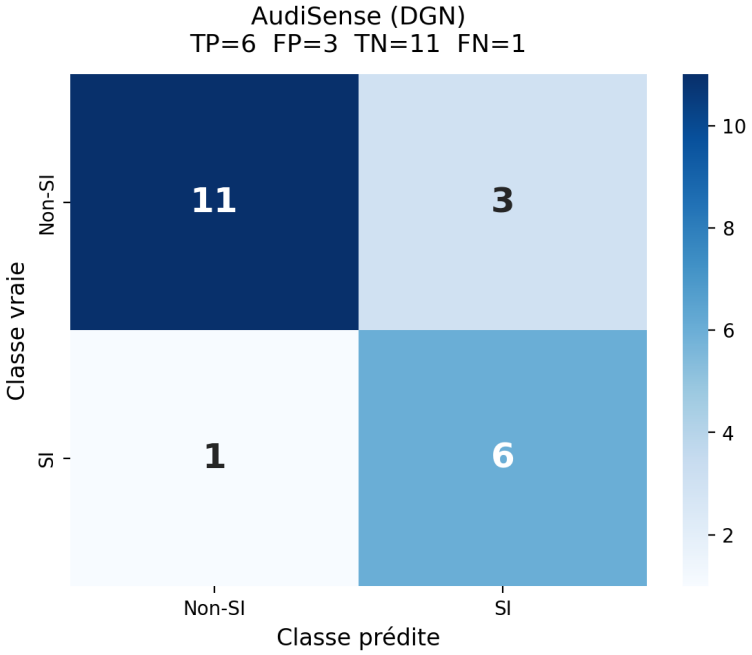

FIGURE 4.5 : Matrice de confusion d'AudiSense sur l'ensemble de test.

Sur les sept vidéos positives, six sont correctement identifiées, soit un rappel de 0,86. Sur les quatorze vidéos négatives, onze sont correctement classées et trois sont prédites positives à tort. La matrice fait donc apparaître un seul faux négatif et trois faux positifs.

L'examen qualitatif des erreurs, conduit sur le texte composite et les compteurs d'engagement des quatre vidéos mal classées, suggère plusieurs facteurs explicatifs. Le faux négatif correspond à une vidéo où l'expression du créateur reste très implicite et où l'engagement, mesuré à un moment donné, n'avait pas encore atteint un volume distinctif. Deux des trois faux positifs concernent des vidéos à fort engagement émotionnel, dont le contenu relève en réalité de la sensibilisation à la santé mentale plutôt que de l'expression d'une idéation suicidaire personnelle. La distinction entre détresse exprimée à la première personne et discours de plaidoyer reste une frontière fine, déjà signalée comme ambiguë dans la littérature (Grant-Allen *et al.*, 2025).

Ces erreurs rappellent que la classification binaire au niveau vidéo agrège des contextes qui, dans la pratique clinique, mériteraient une distinction plus fine. Cette limite intrinsèque de la formulation est à garder à l'esprit lors de toute utilisation opérationnelle.

4.10 SYNTHÈSE COMPARATIVE

Pour résumer cette section, AudiSense obtient les meilleures valeurs ponctuelles sur les quatre métriques principales considérées. Les écarts sont substantiels en termes de F1, de précision, de rappel et de coefficient de Matthews. Trois éléments méritent d'être retenus :

— l'engagement comportemental est, dans ce corpus, le contributeur dominant ;
— le texte multimodal apporte un gain complémentaire à condition d'être combiné de manière adaptative ;
— le portail dynamique module bien sa décision selon la nature de la vidéo, du moins sur l'ensemble de test observé.

4.11 DISCUSSION ET PORTÉE DES RÉSULTATS

Les résultats présentés appellent une lecture prudente. Plusieurs limites doivent être explicitées avant de tirer toute conclusion forte.

La première limite concerne l'évaluation des baselines. Les systèmes inspirés de Mokni & Haoues (2026) et de Liu *et al.* (2022) sont des adaptations à notre corpus, et non des reproductions strictes. Les performances réduites observées pour ces baselines pourraient en partie tenir à cette adaptation, qui les éloigne des conditions originales pour lesquelles elles avaient été conçues. Nous présentons donc ces comparaisons comme indicatives, et non comme une réfutation des méthodes d'origine.

La deuxième limite tient à la nature du corpus. Les 136 vidéos proviennent d'une collecte sur TikTok entre 2024 et 2026, période durant laquelle les pratiques d'algospeak et les codes communautaires peuvent évoluer. La généralisabilité des résultats à d'autres plateformes, à d'autres périodes ou à d'autres communautés linguistiques reste donc une question ouverte, que seules des études complémentaires pourront éclairer.

Malgré ces limites, les tendances observées sont cohérentes entre elles. La supériorité de la branche d'engagement sur la branche textuelle, la complémentarité des deux modalités, et la modulation par vidéo opérée par le portail dynamique pointent dans la même direction. Ces convergences renforcent la plausibilité des hypothèses formulées au Chapitre 1, sans pour autant les démontrer définitivement. Le chapitre suivant approfondit ces considérations en discutant en détail les limitations du travail.

CHAPITRE V
DISCUSSION GÉNÉRALE ET LIMITATIONS

Dans ce chapitre, nous prenons du recul sur les résultats présentés au chapitre précédent et nous discutons en détail les limites du travail. Nous reprenons d'abord les résultats principaux des expérimentations, puis nous examinons successivement les limites méthodologiques, les limites du corpus, les limites de chaque composante de l'architecture, et les considérations éthiques. Cette mise en perspective est essentielle pour situer correctement la portée des conclusions et baliser les pistes d'amélioration.

5.1 PRINCIPAUX RÉSULTATS

Trois résultats principaux ressortent de l'évaluation menée au chapitre 4. Le premier concerne le rôle dominant de la branche d'engagement. Sur notre corpus, l'expert XGBoost appliqué aux quatorze caractéristiques d'engagement atteint déjà un AUPRC de 0,76, contre 0,48 pour l'expert textuel XLM-RoBERTa. Cet écart de près de trente points est trop important pour être attribué au seul hasard, même en tenant compte du faible effectif. Il soutient l'idée que les signaux comportementaux générés par l'auditoire après consommation d'une vidéo encodent une information particulièrement utile dans un environnement où le texte produit par le créateur est dégradé par l'algospeak.

Le deuxième enseignement concerne le gain apporté par la fusion. La combinaison du texte et de l'engagement permet de passer de 0,76 à 0,86 d'AUPRC, soit dix points supplémentaires. Ce gain confirme que le texte, même imparfait, apporte une information complémentaire à celle de l'engagement, à condition que la combinaison soit suffisamment souple pour ne pas pâtir des incompatibilités d'échelle entre les deux espaces de représentation.

Le contraste avec la fusion précoce par perceptron multicouche, qui obtient un AUPRC de seulement 0,29, illustre concrètement le risque associé à une combinaison brutale par concaténation.

Le troisième enseignement concerne le comportement du portail dynamique. Sur l'ensemble de test, le coefficient α moyen pour les vidéos positives est de 0,68, tandis qu'il tombe à 0,36 pour les vidéos négatives. Cette différence indique que le portail ne se contente pas d'apprendre une pondération globale fixe, mais ajuste réellement la balance entre les deux experts en fonction du contenu observé. C'est l'élément le plus encourageant du travail, car il valide empiriquement l'hypothèse architecturale qui sous-tend AudiSense.

5.2 LIMITES MÉTHODOLOGIQUES

5.2.1 TAILLE DE L'ENSEMBLE DE TEST ET SIGNIFICATIVITÉ STATISTIQUE

Comme indiqué au début du chapitre 4, la limite la plus importante du travail tient à la taille de l'ensemble de test. Les estimations de performance sont sensibles à la composition particulière de cet ensemble, ce qui impose de présenter ces résultats comme une étude pilote plutôt que comme une étude confirmatoire.

Cette limite n'est pas le résultat d'un choix arbitraire. Elle est imposée par la nature même de la collecte sur TikTok, où la modération supprime rapidement les contenus sensibles, et par le coût d'une annotation manuelle multi-annotateurs. Néanmoins, elle limite la portée des conclusions que nous pouvons tirer. Une étude confirmatoire devrait s'appuyer sur un corpus de plusieurs centaines de vidéos annotées au minimum, et idéalement de plusieurs milliers, pour permettre une comparaison statistique entre les systèmes.

5.2.2 RISQUE DE SURAPPRENTISSAGE DU PORTAIL

Le réseau de portail dynamique comporte environ une centaine de paramètres apprenables, alors qu'il est entraîné sur le pool combinant entraînement et validation, soit cent quinze exemples. Le rapport entre le nombre de paramètres et le nombre d'exemples est donc proche de l'unité, ce qui invite à la prudence sur la généralité du comportement appris.

Nous avons pris plusieurs précautions pour atténuer ce risque : utilisation d'un taux d'abandon, fonction de perte focale qui concentre l'apprentissage sur les exemples informatifs, et estimation par validation croisée à cinq plis avant l'entraînement final. Nous avons également inclus un empilement logistique parmi les systèmes de référence, afin de pouvoir détecter un éventuel surapprentissage du portail par contraste avec un méta-modèle plus simple. La comparaison n'a pas fait apparaître de gain trompeur, mais elle reste limitée par la taille de l'ensemble de test.

Une étude plus poussée pourrait reproduire l'expérience sur plusieurs ensembles de test indépendants tirés du même corpus, ou simuler des scénarios à plus grande échelle par augmentation contrôlée. Ces analyses dépassent toutefois le périmètre du présent mémoire.

5.2.3 ÉVALUATION DES APPROCHES DE RÉFÉRENCE

Les sept approches de référence ont été implémentées et évaluées selon le même protocole qu'AudiSense, mais elles constituent des adaptations à notre corpus plutôt que des reproductions strictes des travaux originaux. C'est particulièrement le cas pour la variante hybride BERT plus XGBoost inspirée de Mokni & Haoues (2026) et pour la fusion de caractéristiques inspirée de Liu *et al.* (2022), qui avaient été conçues pour des contextes textuels classiques et non pour des vidéos TikTok multilingues. Les performances réduites observées

pour ces baselines pourraient donc tenir, en partie, à la difficulté de transposer fidèlement leurs choix architecturaux à notre cadre.

Nous présentons donc ces comparaisons comme indicatives, et non comme une réfutation des méthodes d'origine. Elles permettent surtout de montrer qu'aucune des approches simples ne suffit, dans notre contexte, à atteindre les performances d'une architecture explicitement conçue pour combiner le texte multilingue et l'engagement comportemental.

5.2.4 CONTRAINTES COMPUTATIONNELLES ET CHOIX D'IMPLÉMENTATION

Comme cela est courant en contexte de ressources limitées, plusieurs choix d'implémentation de ce travail ont été motivés par les ressources de calcul disponibles. Le réglage des hyperparamètres de XGBoost par Optuna a été restreint à quarante essais, ce qui peut avoir laissé inexploré une partie de l'espace de recherche. L'affinage de XLM-RoBERTa a été conduit sur cinq itérations complètes avec une taille de lot de huit exemples, valeurs modestes par rapport aux pratiques courantes en affinage de grands modèles de langage. Une recherche plus large couvrant le taux d'apprentissage, le nombre d'époques et la taille de lot optimale pourrait améliorer les performances de la branche textuelle.

Par ailleurs, le pré-entraînement d'un encodeur multilingue spécialisé en santé mentale, mentionné parmi les pistes futures, n'a pas pu être conduit dans le cadre de ce mémoire en raison du coût computationnel associé. De même, la répétition systématique du protocole sur plusieurs partitions aléatoires différentes de l'ensemble de données, qui aurait permis d'estimer la variabilité associée au choix du découpage, dépasse les ressources qui nous ont été allouées. Ces contraintes ne remettent pas en cause la validité des résultats, mais elles invitent à les interpréter comme un plancher plutôt qu'un plafond pour cette architecture, en

particulier pour la branche textuelle dont les gains potentiels d'un affinage plus poussé restent à confirmer.

5.3 LIMITES DU CORPUS

5.3.1 REPRÉSENTATIVITÉ

Le corpus est constitué de 136 vidéos publiées entre avril 2021 et février 2026, collectées sur TikTok par un seul ensemble de requêtes initiales. Cette collecte ne couvre qu'une fraction du contenu pertinent disponible sur la plateforme, et ses biais ne sont pas pleinement caractérisés. Les codes communautaires, les mots-clics dominants et les pratiques d'algospeak évoluent rapidement sur TikTok, et il est probable que la distribution des contenus dans notre corpus soit différente de celle qu'on observerait avec une collecte effectuée à un autre moment ou par une autre équipe.

La répartition linguistique constitue également une limite. L'anglais est majoritaire dans notre corpus, suivi du français, de l'arabe et du chinois. Cette diversité justifie le recours à XLM-RoBERTa, mais elle est aussi déséquilibrée. Les langues sous-représentées ne bénéficient pas d'un volume d'entraînement suffisant pour qu'on puisse évaluer la performance du système de manière fiable sur chacune d'elles séparément.

5.3.2 SENSIBILITÉ À LA DÉFINITION OPÉRATIONNELLE DE L'IDÉATION SUICIDAIRE

La frontière entre l'expression d'une idéation suicidaire personnelle et un discours de sensibilisation à la santé mentale est par nature floue. Comme nous l'avons souligné au Chapitre 3, le schéma d'annotation que nous avons retenu repose sur l'intention communicative globale de la vidéo, telle qu'elle se manifeste dans son contenu et dans les réactions qu'elle

suscite. Ce critère reste sujet à interprétation, et notre coefficient d'accord inter-annotateurs de 0,76, s'il est substantiel, n'est pas parfait.

L'analyse qualitative des erreurs présentée au chapitre précédent a confirmé cette difficulté. Deux des trois faux positifs d'AudiSense concernent des vidéos relevant en réalité d'une démarche de plaidoyer ou de sensibilisation, et non d'une expression personnelle de détresse. Cette ambiguïté est intrinsèque à la tâche elle-même, et elle limite ce qu'un système automatique peut espérer atteindre, quelle que soit la richesse de son architecture.

5.3.3 ÉVOLUTION TEMPORELLE ET DÉSUÉTUDE

Les pratiques d'algospeak évoluent en réaction aux politiques de modération. Le terme *unalive*, utilisé massivement en 2022 et 2023, fait progressivement place à d'autres codes au fur et à mesure que la plateforme adapte ses filtres. Les caractéristiques d'engagement, elles aussi, sont susceptibles d'évoluer avec les modifications de l'algorithme de recommandation. Un système entraîné sur le corpus actuel pourrait donc voir ses performances se dégrader au fil du temps, à mesure que les codes lexicaux et les comportements d'engagement changent.

Cette désuétude est une caractéristique générale des systèmes de détection sur les médias sociaux. Elle invite à concevoir tout déploiement opérationnel comme un processus continu, avec ré-entraînement périodique sur des données fraîches.

5.4 LIMITES DES COMPOSANTES

5.4.1 LIMITES DE LA TRANSCRIPTION AUDIO

La transcription audio par Whisper a apporté un gain modeste à la branche textuelle, faisant passer son AUPRC de 0,45 à 0,48, soit trois centièmes seulement. Plusieurs facteurs

expliquent cette contribution limitée. Une partie des vidéos ne contient pas de parole intelligible, soit parce qu'elles s'appuient sur de la musique sans paroles ou des effets sonores, soit parce que la voix est couverte par d'autres éléments sonores. Lorsque la parole est présente, sa qualité est souvent dégradée par la compression de la plateforme, par le bruit de fond, ou par des accents et des prononciations qui se prêtent mal à la reconnaissance automatique.

Par ailleurs, nous avons utilisé le mode de traduction de Whisper, qui produit directement une transcription en anglais quelle que soit la langue source. Ce choix simplifie l'intégration au texte composite, mais il introduit un risque de perte d'information lorsque la traduction se révèle approximative. Une étude comparative avec la transcription dans la langue d'origine, suivie d'une analyse multilingue, pourrait éclairer ce point.

5.4.2 LIMITES DE LA BRANCHE D'ENGAGEMENT

Bien que la branche d'engagement soit le contributeur dominant dans notre architecture, elle présente plusieurs limitations. D'abord, les compteurs récoltés sont des valeurs ponctuelles, mesurées au moment de la collecte. Or les vidéos accumulent leur engagement progressivement, parfois sur des semaines, et la dynamique de cette accumulation pourrait porter une information distincte de la valeur cumulée. Notre approche statique perd cette dimension temporelle.

Ensuite, la sensibilité de l'engagement à l'algorithme de recommandation de TikTok introduit un biais potentiel. Une vidéo très exposée par l'algorithme accumulera mécaniquement plus d'interactions, qu'elle exprime ou non une idéation suicidaire. Les caractéristiques relatives, comme les ratios d'interaction, atténuent en partie ce biais en normalisant par le nombre de vues, mais elles ne le suppriment pas entièrement.

Enfin, la branche d'engagement, par construction, ne peut produire de prédiction utile que pour des vidéos ayant déjà accumulé un volume d'interactions suffisant. Pour une vidéo récemment publiée et peu vue, les compteurs ne sont pas informatifs, et c'est sur la branche textuelle qu'il faudrait alors s'appuyer. Cette dépendance à l'ancienneté de la vidéo est un point qu'une éventuelle mise en service opérationnelle devrait considérer attentivement.

5.4.3 ALGOSPEAK NON QUANTIFIÉ

Notre travail repose en grande partie sur l'argument que l'algospeak dégrade le signal textuel. Cette idée est étayée par la littérature et par l'observation qualitative du corpus, mais nous n'avons pas, mesuré directement l'ampleur de cette dégradation. Nous ne disposons pas, par exemple, d'un score qui quantifierait, pour chaque vidéo, le degré d'obfuscation lexicale. Une telle mesure permettrait de valider plus rigoureusement notre interprétation, en montrant que l'écart de performance entre la branche textuelle et la branche d'engagement est plus marqué sur les vidéos fortement obfusquées.

Construire un score d'algospeak n'est toutefois pas trivial. Il faudrait au minimum disposer d'un lexique de référence des codes connus, et idéalement d'une mesure de la divergence entre le texte observé et un texte non obfusqué de même contenu. C'est une piste de recherche en soi, que nous mentionnons au chapitre suivant.

5.5 CONSIDÉRATIONS ÉTHIQUES APPROFONDIES

5.5.1 CONDITIONS DE LA COLLECTE

Toutes les données utilisées proviennent de vidéos publiquement accessibles, sans interaction directe avec les utilisateurs et sans sollicitation de données privées. Selon l'article 2.2 de l'*Énoncé de politique des trois Conseils : Éthique de la recherche avec des êtres*

humains (Conseil de recherches en sciences humaines du Canada *et al.*, 2022), l'analyse de tels contenus ne requiert pas d'approbation préalable d'un comité d'éthique. Nous avons par ailleurs pris des précautions de minimisation : les identifiants d'utilisateurs ne sont pas utilisés par les modèles, et les exemples textuels cités à titre illustratif ont été paraphrasés.

Ces précautions, conformes aux pratiques recommandées dans la littérature, n'épuisent toutefois pas les questions éthiques que soulève le travail. La frontière entre données publiques et données sensibles est particulièrement ténue lorsque les contenus touchent à la santé mentale. Une personne qui partage publiquement une vidéo exprimant sa détresse ne consent pas explicitement à ce que cette vidéo serve à entraîner un système de classification. Nous estimons cependant que l'intérêt scientifique de cette recherche, ainsi que les précautions de non-identification que nous avons prises, justifient la démarche dans le cadre académique présent. Un déploiement opérationnel à plus grande échelle exigerait une réflexion éthique additionnelle.

5.5.2 RISQUES D'USAGE DU SYSTÈME

Un système automatique de détection d'idéation suicidaire peut être employé à des fins très variées, dont certaines sont bénéfiques et d'autres potentiellement nuisibles. Du côté bénéfique, on peut imaginer un dispositif d'alerte précoce intégré à une politique de prévention, qui orienterait des modérateurs humains vers les contenus à risque pour qu'ils proposent des ressources d'aide à leurs auteurs. C'est l'usage que nous estimons légitime, à condition que la décision finale d'intervention reste humaine.

Du côté des risques, plusieurs scénarios méritent d'être anticipés. Un système déployé sans accompagnement humain pourrait produire des faux positifs, susceptibles de stigmatiser des utilisateurs qui n'expriment pas de détresse réelle, ou des faux négatifs qui passeraient à

côté de cas authentiques. Le rappel de notre système, bien que relativement élevé, n'atteint pas la perfection, et nos faux positifs touchent justement les discours de plaidoyer dont la stigmatisation serait particulièrement contre-productive.

Un risque distinct concerne la possibilité que le système soit détourné par des acteurs malveillants, par exemple pour repérer et cibler des personnes vulnérables. Ce risque est consubstantiel à toute publication ouverte de systèmes de détection en santé mentale, et il invite à mettre en garde clairement les utilisateurs potentiels contre les usages détournés.

5.5.3 POSITION DE PRUDENCE

Pour ces raisons, nous considérons que notre travail ne constitue pas et ne saurait constituer un outil de diagnostic clinique. Il s'agit d'une recherche exploratoire dont les sorties devraient, dans toute application opérationnelle envisageable, être systématiquement intégrées à un dispositif de triage humain, sous la supervision de professionnels qualifiés en santé mentale. Toute utilisation directe d'un système automatique de ce type, sans cette médiation, serait à nos yeux contraire à l'esprit dans lequel ce travail a été conduit.

5.6 TRAVAUX FUTURS

Plusieurs pistes méritent d'être explorées dans la continuité de ce travail. Nous les présentons brièvement, sans prétendre à l'exhaustivité.

La priorité revient à l'extension du corpus. Une étude confirmatoire, capable de produire des comparaisons statistiquement significatives entre systèmes, supposerait plusieurs centaines de vidéos annotées, idéalement collectées de manière continue pour compenser la suppression rapide des contenus par la modération.

Une deuxième direction concerne la dynamique temporelle de l'engagement. Nos compteurs sont des valeurs statiques, mesurées à un instant donné ; or la vitesse d'accumulation des interactions porte vraisemblablement une information distincte de leur valeur cumulée. Une collecte longitudinale, ou l'ajout de caractéristiques décrivant la trajectoire d'engagement, permettrait de capter cette dimension.

Une troisième piste consisterait à quantifier directement l'algospeak. Notre travail postule que l'obfuscation lexicale dégrade le signal textuel, sans le mesurer explicitement. Un score d'obfuscation, fondé sur un lexique de codes connus ou sur une divergence statistique avec un texte de référence, permettrait de stratifier l'évaluation selon le degré de dégradation et de conditionner explicitement le portail dynamique sur cette information.

Sur le plan de l'architecture, le recours à des encodeurs spécialisés en santé mentale, multilingues, et l'évaluation de grands modèles de langage généralistes en peu-shot, constituent deux pistes prometteuses pour gagner en performance sans alourdir la collecte d'annotations.

Enfin, la validation transplate-forme sur Reddit, X ou Instagram, et la réflexion sur un déploiement responsable, intégrant un dispositif de triage humain et une évaluation par sous-populations, dépassent le cadre purement technique du présent travail et appellent une collaboration interdisciplinaire avec des cliniciens et organismes de prévention.

5.7 SYNTHÈSE DU CHAPITRE

Ce chapitre a passé en revue les principales limites du travail. La taille restreinte de l'ensemble de test, la composition particulière du corpus, les limites propres à la transcription audio et à l'engagement, l'absence d'une quantification directe de l'algospeak, et les enjeux éthiques d'un déploiement opérationnel constituent autant de réserves qui invitent à interpréter les résultats avec prudence. Aucune de ces limites ne remet en cause l'intérêt de la démarche,

mais toutes appellent à la consolidation par des travaux complémentaires, dont les principales pistes ont été esquissées à la section précédente. La conclusion générale synthétise désormais les apports du travail et ouvre sur ses prolongements possibles.

CONCLUSION GÉNÉRALE

Ce mémoire est parti d'un constat simple : sur TikTok, le texte ne suffit plus. Les créateurs déforment volontairement les mots sensibles pour passer entre les mailles de la modération, et les corpus annotés disponibles restent minuscules. Travailler sur cette plateforme avec les méthodes habituelles du traitement du langage revient donc à se priver, dès le départ, d'une partie de l'information.

Nous avons donc essayé autre chose. AudiSense combine un encodeur multilingue XLM-RoBERTa pour le texte et un classifieur XGBoost appliqué à quatorze caractéristiques d'engagement, et laisse à un petit réseau de portail dynamique le soin de décider, vidéo par vidéo, sur lequel des deux experts s'appuyer davantage. Le système a été évalué sur un corpus que nous avons construit nous-mêmes : 136 vidéos TikTok et 1 339 commentaires, annotés à la main avec un accord inter-annotateurs $\kappa = 0{,}76$.

Les chiffres parlent d'eux-mêmes. AudiSense atteint un AUPRC de 0,86 et un ROC-AUC de 0,92 sur l'ensemble de test, en tête des neuf systèmes comparés. Mais c'est l'étude d'ablation qui est sans doute la plus parlante : prise seule, la branche d'engagement obtient 0,76 d'AUPRC, contre 0,48 pour la branche textuelle. Autrement dit, quand l'algospeak dégrade le texte, l'engagement reste un signal solide. La fusion des deux modalités fait monter le score à 0,86, ce qui montre que le texte garde son utilité, à condition qu'on sache quand lui faire confiance. C'est justement ce que fait le portail dynamique : sur les vidéos positives, il attribue en moyenne un poids de 0,68 au texte, contre 0,36 pour les vidéos négatives. Il ne se contente donc pas d'apprendre une pondération globale, il s'adapte vraiment au contenu de chaque vidéo.

Cela dit, nous ne voulons pas surinterpréter ces résultats. L'ensemble de test ne contient que 21 vidéos, dont 7 positives. Cela suffit pour dégager une tendance et avancer une hypothèse,

mais pas pour parler de significativité statistique au sens strict. Nous présentons donc ce travail comme une étude pilote, à confirmer sur un corpus plusieurs fois plus grand. D'autres limites s'ajoutent : la transcription audio par Whisper ne change pas grand-chose, parce que beaucoup de vidéos contiennent surtout de la musique ou du bruit ; les compteurs d'engagement sont ponctuels et nous ne captons pas la dynamique de leur accumulation ; et l'algospeak n'est jamais quantifié directement, donc notre lecture du phénomène reste en partie qualitative.

La suite naturelle de ce travail est claire. Il faut d'abord agrandir le corpus, parce que c'est ce qui conditionne tout le reste. Il faudrait aussi construire un score d'obfuscation pour mesurer l'algospeak directement, et non seulement le supposer. Sur le plan de la modélisation, des encodeurs spécialisés en santé mentale, multilingues, et l'évaluation de grands modèles de langage en peu-shot pourraient apporter des gains supplémentaires. Enfin, tout déploiement réel devrait passer par un dispositif de triage humain, conçu avec des cliniciens, et non se substituer à eux.

Un système comme AudiSense n'a pas vocation à remplacer un professionnel de santé mentale. Au mieux, il peut aider à repérer plus tôt des contenus qui mériteraient une attention humaine. C'est dans ce cadre, et seulement dans ce cadre, que ce travail prend son sens.

BIBLIOGRAPHIE

Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. (2019). Optuna : A Next-Generation Hyperparameter Optimization Framework. Dans *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631.

Baltrusaitis, T., Ahuja, C. & Morency, L.-P. (2019). Multimodal machine learning : A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), 423–443.

Basch, C. H., Donelle, L., Fera, J. & Jaime, C. (2022). Deconstructing TikTok videos on mental health : cross-sectional, descriptive content analysis. *JMIR formative research*, *6*(5), e38340.

Chen, T. & Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. Dans *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Chicco, D. & Jurman, G. (2020). The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, *21*, 6. doi: 10.1186/s12864-019-6413-7

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Dans *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8440–8451.

Conseil de recherches en sciences humaines du Canada, Conseil de recherches en sciences naturelles et en génie du Canada & Instituts de recherche en santé du Canada (2022, décembre). *Énoncé de politique des trois Conseils : Éthique de la recherche avec des êtres humains (EPTC 2)*. Ottawa. Secrétariat sur la conduite responsable de la recherche. Repéré à https://ethics.gc.ca/fra/policy-politique_tcps2-eptc2_2022.html

Coppersmith, G., Dredze, M. & Harman, C. (2014). Quantifying mental health signals in Twitter. Dans *Proceedings of the workshop on computational linguistics and clinical psychology : From linguistic signal to clinical reality*, pp. 51–60.

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 215–242.

Cui, J., Zhang, X. & Wang, L. (2024). Speech-Based Suicide Risk Assessment Using Whisper and Large Language Models. Dans *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 3456–3460.

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.

Ghanadian, H., Nejadgholi, I. & Al Osman, H. (2025). Improving Suicidal Ideation Detection in Social Media Posts : Topic Modeling and Synthetic Data Augmentation Approach. *JMIR Formative Research*, *9*, e63272. doi: 10.2196/63272

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J. & Dutta, R. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific reports*, *7*(1), 45141.

Grant-Allen, G., Wang, L., Amini, J., Dhaliwal, S., Sinyor, M. & Mitchell, R. H. (2025). Self-Harm and Suicide-Related Content on TikTok : Thematic Analysis. *Journal of Medical Internet Research*, *27*, e77828.

Hanley, J. A. & McNeil, B. J. (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, *143*(1), 29–36.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P. & Cambria, E. (2022). MentalBERT : Publicly Available Pretrained Language Models for Mental Healthcare. pp. 7184–7190.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017). Focal Loss for Dense Object Detection. Dans *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988.

Liu, Z., Hu, B., Wang, S., Lian, W. & Zhong, Y. (2022). Suicide Risk Assessment on Weibo : A Multi-Source Feature Fusion Approach. *IEEE Transactions on Affective Computing*, *13*(2), 1036–1046.

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

Luceri, L., Cardoso, F. & Ferrara, E. (2025). Behavioral Engagement Signatures as Manipulation-Resistant Signals in Social Media Inference. *Social Networks*, *80*, 100–112.

Mokni, R. & Haoues, M. (2026). Hybrid BERT-XGBoost transformer models for predicting mental health from social media. *World Wide Web*, *29*(2), 21.

Nasiri, A., Yusefzadeh, H., Amerzadeh, M., Moosavi, S. & Kalhor, R. (2022). Measuring inequality in the distribution of health human resources using the Hirschman-Herfindahl index : a case study of Qazvin Province. *BMC health services research*, *22*(1), 1161.

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*.

Pastro, M., Ribeiro, F. H. & Almeida, V. (2026). Detecting Suicidal Ideation on TikTok : Challenges and Baselines. *EPJ Data Science*, *15*(1), 4. doi: 10.1140/epjds/s13688-026-00300-5

Pennington, J., Socher, R. & Manning, C. D. (2014). GloVe : Global Vectors for Word Representation. Dans *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the International Conference on Machine Learning (ICML)*, *202*, 28492–28518.

Saerens, M., Latinne, P. & Decaestecker, C. (2002). Adjusting the Outputs of a Classifier to New a Priori Probabilities : A Simple Procedure. *Neural Computation*, *14*(1), 21–41.

Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Sawhney, R., Joshi, H., Flek, L. & Shah, R. (2021). Phase : Learning emotional phase-aware representations for suicide ideation detection on social media. Dans *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics : main volume*, pp. 2415–2428.

Sawhney, R., Joshi, H., Gandhi, S. & Shah, R. R. (2021). Towards ordinal suicide ideation detection on social media. Dans *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 22–30.

Sawhney, R., Joshi, H., Shah, R. & Flek, L. (2021). Suicide ideation detection via social and temporal user representations using hyperbolic learning. Dans *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics : human language technologies*, pp. 2176–2190.

Shaikh, N. (2025). Multimodal Harmful Content Detection Using BERT, Visual, and Audio Representations. Dans *Proceedings of the ACM International Conference on Multimedia*, pp. 2341–2349.

Spangenberg, L., Teismann, T. & Höller, I. (2025). Thinking About Suicide for a Long Time : A Scoping Review of Empirical Studies on Persistent Suicidal Ideation. *Suicide and Life-Threatening Behavior*, *55*(6), e70070. doi: 10.1111/sltb.70070

Steen, E., Yurechko, K. & Klug, D. (2023). You can (not) say what you want : Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media+ Society*, *9*(3), 20563051231194586.

Sun, W., Yin, H., Bai, J. & Chen, J. (2025). Dynamic Fusion Multimodal Network for SpeechWellness Detection. Dans *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2002–2007. IEEE.

Wolpert, D. H. (1992). Stacked Generalization. *Neural Networks*, *5*(2), 241–259.

Xue, R. & Marculescu, R. (2022). DynMM : A Dynamic Multimodal Fusion Method for Vision and Language. Dans *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 28345–28358.

Zhang, X., Li, M. & Chen, W. (2025). KETCH : Knowledge-Enhanced Transformer for Clinical Mental Health Reasoning. Dans *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 14201–14209.