Une Approche Unifiée pour la Détection du Sexisme dans les Mèmes des Réseaux Sociaux dans des Contextes d'Evaluation Stricts et Souples

Prénom1 Nom1[1] Prénom2 Nom2[1]
(1) Lab, adresse, CP Ville, Pays
utrucmuche@lab.fr, umachinchose@adresse-academique.be

RÉSUMÉ

La détection automatique du sexisme dans les mèmes se heurte à la nature multimodale des contenus et à la pluralité des perspectives portées par les annotateurs. Si les approches classiques visent un consensus souvent artificiel, le défi EXIST 2025 souligne la nécessité de concilier évaluations strictes (binaires) et souples (probabilistes). Nous proposons une architecture unifiée modélisant conjointement ces deux paradigmes. Notre approche repose sur une stratégie d'ensemble entraînée sur des partitions asymétriques pour capturer la distribution des avis, couplée à un apprentissage supervisé ciblé sur les cas limites. Les expérimentations menées sur le corpus EXIST Meme démontrent l'efficacité de cette synergie, avec une progression de +7,73% sur la métrique souple et +26,47% en évaluation stricte par rapport à l'état de l'art. Cette approche souligne l'efficacité d'un cadre unifié combinant inférence déterministe et estimation probabiliste, offrant ainsi une réponse robuste à l'ambiguïté inhérente à l'interprétation des mèmes.

ABSTRACT

A Unified Approach for Sexism Detection in Social Media Memes Under Hard and Soft Evaluation Settings

The automatic detection of sexism in memes faces the multimodal nature of the content and the plurality of perspectives held by annotators. While traditional approaches aim for a consensus that is often artificial, the EXIST 2025 challenge highlights the need to reconcile hard (binary) and soft (probabilistic) evaluations. We propose a unified architecture that jointly models these two paradigms. Our approach is based on an ensemble strategy trained on asymmetric partitions to capture the distribution of opinions, combined with supervised learning focused on borderline cases. Experiments conducted on the EXIST Meme corpus demonstrate the effectiveness of this synergy, with an improvement of +7.73% on the soft metric and +26.47% in the hard evaluation compared to the state of the art. This approach underscores the effectiveness of a unified framework combining deterministic inference and probabilistic estimation, thereby offering a robust response to the ambiguity inherent in meme interpretation.

MOTS-CLÉS : Identification du Sexisme, Classification de Texte, Classification d'Images.

KEYWORDS: Sexism Identification, Text Classification, Image Classification.

ARTICLE SOUMIS À : The 26th International Conference of Web Engineering (ICWE).

1 Introduction

La montée de l'harcèlement en ligne, notamment des contenus sexistes et misogynes sur les réseaux sociaux, a renforcé le besoin de systèmes fiables de détection automatique. Parmi les récentes initiatives de benchmarking, le défi EXIST (Plaza *et al.*, 2025) s'est imposé comme une tâche de référence pour évaluer les modèles d'identification du sexisme à travers plusieurs modalités, notamment les tweets, les mèmes et les vidéos. L'un des aspects distinctifs d'EXIST est son double paradigme d'évaluation : les systèmes sont évalués à la fois selon des **étiquettes hard (strictes)**, reflétant des attributions de classe déterministes, et selon des **étiquettes soft (souples)**, reflétant la distribution des jugements de plusieurs annotateurs. L'évaluation souple est particulièrement pertinente pour des phénomènes où la perception individuelle engendre des désaccords intrinsèques, à l'instar de la compréhension du sexisme. À notre connaissance, EXIST est le premier benchmark à aborder la détection du sexisme dans des contextes d'évaluation à la fois strict et souple.

Malgré les progrès réalisés dans les architectures multimodales et basées sur des transformateurs pour l'identification du sexisme dans les mèmes, les solutions existantes soumises au défi EXIST révèlent systématiquement un écart de performance entre les deux contextes d'évaluation. Les modèles qui obtiennent de bonnes performances dans le cadre de mesures d'évaluation strictes ne parviennent souvent pas à saisir avec précision les distributions de probabilité de six annotateurs requises pour l'évaluation souple, et vice versa (Plaza *et al.*, 2025). Ces divergences suggèrent que les approches actuelles ne sont pas entièrement équipées pour concilier la différence entre la classification déterministe et la modélisation d'une interprétation nuancée.

Ces observations soulignent la nécessité de développer des architectures capables de réconcilier, au sein d'un cadre unifié, les paradigmes d'évaluation stricts et souples. Dans ce travail, nous relevons ce défi par une approche articulée autour de deux axes :

— **Un apprentissage par diversité de partitions :** à partir des distributions d'annotations souples (*soft labels*), nous exploitons une **architecture d'ensemble légère**. Celle-ci repose sur deux modèles entraînés sur des partitions de données délibérément asymétriques, permettant de capturer la variance des opinions sans recourir à des connaissances exogènes (ex : métadonnées ou agents externes).

— **Un arbitrage par modèle de vision-langage (VLM) :** pour les contenus où l'accord entre annotateurs est faiblement polarisé, nous introduisons une étape de résolution par un VLM. Contrairement aux approches probabilistes continues, le VLM est sollicité pour produire une décision catégorielle explicite. En s'appuyant sur ses capacités de raisonnement contextuel, il agit comme un mécanisme de discrétisation forcée, permettant de trancher les cas d'incertitude là où les modèles d'ensemble ne parviennent pas à dégager un consensus statistique net.

Nos résultats démontrent que la stratégie de partitionnement déséquilibré proposée permet d'obtenir des performances de pointe en matière d'évaluation souple grâce à une architecture légère. De plus, l'intégration de la classification binaire uniquement sur les cas limites comble efficacement le fossé entre les deux paramètres d'évaluation, introduisant un système robuste de détection du sexisme qui reflète mieux l'ambiguïté et la multiplicité du jugement humain sur les mèmes des réseaux sociaux.

2 Travaux Connexes

TIB-VA à SemEval-2022 Task 5 (Hakimov *et al.*, 2022) a défini une base de référence parmi les systèmes de détection de la misogynie et du sexisme, en tirant parti d'une architecture multimodale à fusion précoce (early fusion) pour combiner des caractéristiques textuelles et visuelles afin de détecter et de classer les mèmes misogynes, montrant que l'intégration des modalités au stade de l'entrée améliore la reconnaissance des contenus haineux fondés sur des éléments visuels. Cette approche a ensuite été adaptée et proposée par l'équipe NICA lors de l'EXIST 2024 (Naebzadeh *et al.*, 2024).

En utilisant la concaténation de divers modèles sur des signaux textuels uniquement, Victor-UNED lors de l'EXIST 2024 (Ruiz *et al.*, 2024) a présenté une nouvelle stratégie de prédiction des étiquettes probabilistes. Au départ, les modèles sont utilisés pour identifier les mèmes présentant un niveau d'accord plus élevé. Une fois ces instances sélectionnées, un autre modèle génère des prédictions pour celles qui présentent un niveau d'accord plus faible. Parmi les systèmes multimodaux qui ont soumis leurs résultats dans la même édition du Challenge, UMUTeam (Pan *et al.*, 2025) a exploité des encodeurs de texte et d'images pour générer et fusionner des caractéristiques multimodales utilisées pour prédire à la fois les étiquettes dures et douces. TrankilTwice à EXIST 2025 (Italiani *et al.*, 2025) propose un système multimodal de détection du sexisme pour les mèmes qui intègre des connaissances externes via le sous-titrage de mèmes basé sur VLM, enrichissant les représentations textuelles afin de mieux saisir le sexisme implicite et visuellement ancré et d'améliorer la compréhension intermodale, atteignant des résultats de pointe dans le cadre du défi EXIST dans le paradigme de l'évaluation souple.

Bien que ces méthodes soient très performantes, elles se caractérisent souvent par une baisse de performance entre les paramètres d'évaluation stricte et souple (Carrillo-Casado *et al.*, 2024; Italiani *et al.*, 2025; Ruiz *et al.*, 2024), ainsi que par une dégradation des performances sur les instances non anglaises, ce qui limite la généralisation des approches proposées.

Récemment, Qwen3Guard (Zhao *et al.*, 2025) a amélioré la détection des discours haineux en ajoutant une étiquette « Controversé » pour les cas ambigus, en montrant que le déséquilibre des étiquettes affecte les décisions du modèle et en utilisant le vote d'ensemble sur des ensembles de données différemment déséquilibrés afin de mieux identifier les contenus clairs et ambigus.

3 Aperçu des Tâches et des Données

Ce travail se concentre sur la tâche 2.1 du défi EXIST 2025 (Plaza *et al.*, 2025), qui consiste à détecter automatiquement le sexisme dans les mèmes. Plus précisément, les systèmes doivent déterminer si un mème donné exprime des idées sexistes ; en véhiculant des croyances ou des stéréotypes sexistes, en décrivant une situation ou un comportement sexiste, ou en critiquant des idées ou des comportements sexistes.

Le jeu de données EXIST 2025 Memes est composé d'images de mèmes en anglais et en espagnol annotées pour la détection du sexisme en ligne. Chaque mème du jeu de données est annoté par six annotateurs différents, de sorte que :

— pour l'évaluation stricte, la vérité terrain est déterminée par un vote à la majorité. Seulement les données pour lesquelles il existe un consensus majoritaire entre les annotateurs sont utilisées pour l'évaluation.

— pour l'évaluation souple, la vérité terrain correspond à la distribution des réponses des annotateurs. Par conséquent, dans cette phase, toutes les données sont utilisées pour l'évaluation.

Langue	Training	Test
Anglais	2010	513
Espagnol	2034	540
Tout	4044	1053

TABLE 1 – Répartition par langue.

Classe	Proportion (%)	Sous-classe	Proportion (%)
Sexiste	50.39	[6, 0]	14.34
		[5, 1]	19.24
		[4, 2]	16.82
Non sexiste	34.17	[0, 6]	9.32
		[1, 5]	11.44
		[2, 4]	13.40
Non étiqueté	15.44	[3, 3]	15.44

TABLE 2 – Répartition des réponses des annotateurs (EXIST 2025 Mèmes).

Le Tableau 1 montre la répartition linguistique des mèmes dans l'ensemble de données, tandis que le Tableau 2 présente la répartition des étiquettes dures et des étiquettes souples parmi les données prises en compte. La classe « Non étiqueté » correspond aux enregistrements controversés pour lesquels les annotateurs ne sont pas parvenus à un consensus.

4 Mesures d'Evaluation

La métrique d'évaluation officielle pour ce défi est la mesure de contraste d'information (ICM) (Amigo & Delgado, 2022), une métrique d'évaluation hiérarchique qui compare les étiquettes prédites et les étiquettes de référence tout en tenant compte des distances sémantiques entre les classes dans la hiérarchie. Cela la rend particulièrement adaptée aux tâches de classification à étiquettes rigides impliquant des sous-classes très détaillées. Nous adoptons également l'ICM-Soft, une extension de l'ICM au cadre des étiquettes souples, qui évalue les distributions de probabilité prédites par rapport à celles fournies par les annotateurs et récompense les modèles qui capturent l'incertitude et le désaccord des annotateurs (Mostafazadeh Davani *et al.*, 2022). Dans ce rapport, nous examinerons en outre leurs variantes normalisées ICM Norm et ICM-Soft Norm, prenant des valeurs comprises entre 0 et 1, afin de fournir une évaluation plus interprétable des performances du modèle. Pour les résultats de classification binaire, le score F1 est rapporté, tandis que la métrique d'entropie croisée (Goodfellow *et al.*, 2016) est utilisée pour évaluer la prédiction de l'accord entre annotateurs.

5 Partitionnement des Données

Comme le montre la Figure 1, à partir de l'ensemble complet de données étiquetées (4044 enregistrements), l'ensemble de données a été divisé en trois partitions 80/10/10, correspondant aux ensembles d'entraînement, de validation et de test. En particulier :

— Dans le cadre souple, l'ensemble d'apprentissage a été divisé en deux partitions non superposables, nommées **Partie 1** et **Partie 2**. Chaque partition contient le même nombre de mèmes

indécis, pour lesquels aucun accord majoritaire n'a été atteint parmi les annotateurs. Cependant, ces deux partitions sont **déséquilibrées** quant au nombre d'enregistrements sexistes et non sexistes qu'elles contiennent. Plus précisément, la partie 1 contient 75 % des mèmes sexistes de l'ensemble d'apprentissage et 25 % des instances non sexistes du même ensemble, tandis que la partie 2 contient 75 % des mèmes non sexistes et 25 % des mèmes sexistes de l'ensemble d'apprentissage. Les parties 1 et 2 sont ensuite utilisées pour entraîner deux modèles distincts, appelés **Modèle 1** et **Modèle 2**, qui sont tous deux utilisés pour générer des prédictions sur les ensembles de validation et de test dans un paradigme d'ensemble.

— L'ensemble **Validation** a été utilisé pour affiner les hyperparamètres d'entraînement (taux d'apprentissage, nombre d'époques, taille des lots) et pour tester la meilleure stratégie de combinaison de prédictions à partir des deux modèles.

— L'ensemble **Test** a finalement été utilisé pour évaluer les performances de l'architecture ensembliste sur des données jamais vues auparavant.

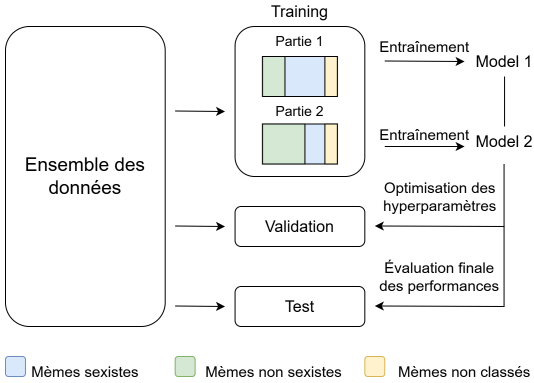

FIGURE 1 – Stratégie de partitionnement des données dans laquelle deux modèles sont entraînés sur des partitions distinctes et déséquilibrées en termes de classes. Le Modèle 1 et le Modèle 2 partagent la même architecture, illustrée à la Figure 2

6 Méthodologie

L'architecture globale, structurée autour d'un mécanisme d'attention croisée (*cross-modal*) pour fusionner les représentations visuelles et textuelles, est détaillée dans la Figure 2.

Comme présenté dans la Section 5, deux modèles sont entraînés sur des partitions distinctes de l'ensemble d'entraînement. Bien que les données sur lesquelles ils sont entraînés diffèrent, le Modèle 1 et le Modèle 2 partagent la même architecture, qui sera présentée en détail dans cette section.

Chaque mème M se compose d'une composante visuelle I, où $I \in R^{H \times L \times C}$ désigne l'image brute représentée par sa hauteur, sa largeur et ses canaux de couleur, ainsi que d'une composante textuelle $T = t_1, t_2, \ldots, t_n$, correspondant à la séquence de mots du texte superposé. Les entrées bimodales sont traitées séparément par un **Encodeur intermodal** (*Cross-modal Embedder*) :

$$\mathbf{v} = f_v(I), \quad \mathbf{u} = f_t(T)$$

où f_v et f_t sont l'**Encodeur d'image** et l'**Encodeur de texte**, produisant une représentation de l'image $\mathbf{v} \in R^d$ et une représentation du texte $\mathbf{u} \in R^d$ encodées dans un espace latent partagé de dimension d.

Afin d'enrichir chaque modalité avec des informations contextuelles provenant de l'autre, un mécanisme d'**attention intermodale bidirectionnelle** est appliqué. Des connexions résiduelles sont intégrées afin de préserver les caractéristiques propres à chaque modalité tout en incorporant des informations complémentaires, ce qui conduit à des représentations enrichies de l'image et du texte comme suit :

— **Attention Image-vers-Texte :**

$$\mathbf{v}^* = \mathbf{v}' + \mathrm{Attn}(\mathbf{v}', \mathbf{u}', \mathbf{u}') \tag{1}$$

où la représentation d'image $\mathbf{v}'$ porte son attention sur la représentation textuelle $\mathbf{u}'$ produisant une représentation enrichie $\mathbf{v}^*$ intégrant la sémantique du texte.

— **Attention Texte-vers-Image :**

$$\mathbf{u}^* = \mathbf{u}' + \mathrm{Attn}(\mathbf{u}', \mathbf{v}', \mathbf{v}') \tag{2}$$

où la représentation textuelle $\mathbf{u}'$ porte son attention sur la représentation d'image $\mathbf{v}'$, produisant une représentation textuelle enrichie $\mathbf{u}^*$ reflétant les indices visuels.

Cette interaction bidirectionnelle garantit que chaque modalité est informée contextuellement par l'autre, conduisant à une représentation plus complète :

$$\mathbf{h} = \mathbf{v}^* \oplus \mathbf{u}^* \tag{3}$$

où $\oplus$ désigne la concaténation de vecteurs.

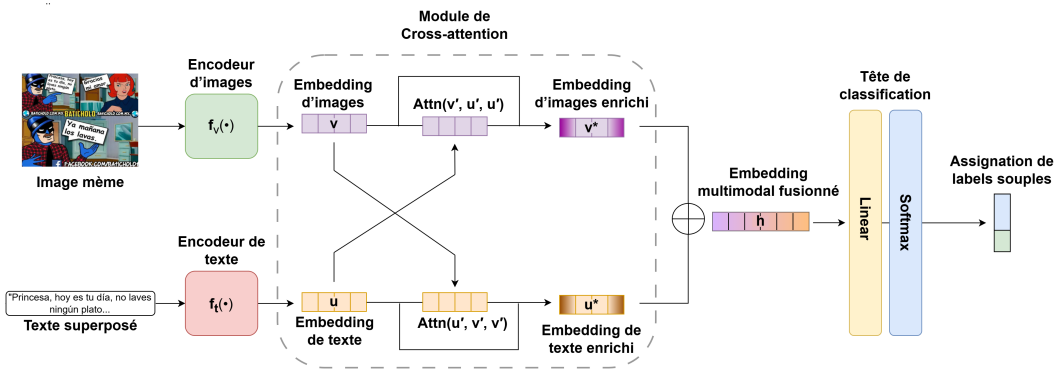

FIGURE 2 – Architecture du Modèle 1 et du Modèle 2 pour la prédiction à étiquettes souples dans la tâche d'identification du sexisme. Bien que les deux modèles reposent sur le même pipeline, ils sont entraînés à partir de partitions de données différentes et déséquilibrées en termes de classes.

Les étapes précédentes sont effectuées pour chaque mème, créant ainsi une représentation vectorielle unifiée pour chaque élément du jeu de données. L'ensemble des représentations des mèmes est défini comme suit :

$$\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\},$$

où N est le nombre de mèmes dans le jeu de données et $\mathbf{h}_i \in R^d$ est la représentation du mème i, obtenue par la concaténation des représentations produites par le module d'attention intermodale.

Enfin, les représentations unifiées des mèmes sont transmises à un perceptron multicouche (MLP) servant de **tête de classification** afin de produire des étiquettes probabilistes. Le réseau comprend une couche de projection d'entrée, une transformation non linéaire via une activation ReLU, ainsi qu'une couche de sortie produisant un vecteur de logits bidimensionnel représentant le niveau de désaccord entre les annotateurs.

Le modèle est entraîné en minimisant la divergence de Kullback–Leibler (KL) entre la distribution prédite $\hat{y}$ et la distribution réelle y :

$$\mathcal{L}_{\mathrm{KL}}(y, \hat{y}) = \sum_i y_i \log\left(\frac{y_i}{\hat{y}_i}\right) \tag{4}$$

Étant donné que les cibles de cette étude sont des distributions de probabilité souples, l'objectif est d'ajuster l'ensemble de la distribution plutôt que la classe la plus probable. L'entropie croisée est conçue pour des étiquettes one-hot, tandis que la divergence KL mesure directement l'écart entre les distributions prédite et cible, ce qui la rend plus appropriée pour l'apprentissage à partir d'étiquettes souples.

Comme présenté dans la Section 5, durant l'entraînement du modèle, l'ensemble de validation est utilisé pour ajuster plusieurs hyperparamètres, tels que le taux d'apprentissage, le nombre d'époques et la taille des lots (batch size). Chaque mème est traité indépendamment par les deux modèles, produisant une prédiction individuelle chacun. Étant donné les deux vecteurs de sortie issus des deux modèles, nous les combinons par une multiplication élément par élément suivie d'une normalisation L1.

Cette interaction multiplicative atténue les prédictions faiblement soutenues par les deux modèles tout en amplifiant les preuves conjointement fortes, entraînant une réduction de l'entropie principalement lorsque les deux prédictions sont alignées. Contrairement à une moyenne linéaire, cette opération valorise le consensus et empêche une seule prédiction de dominer. La normalisation L1 garantit que la distribution résultante appartient au simplexe des probabilités et préserve les rapports de vraisemblance relatifs, rendant la distribution combinée adaptée à l'inférence du degré d'accord des annotateurs.

Lorsque la différence entre les probabilités prédites des deux classes se réduit, l'intégration de contextes externes ou culturels qui font défaut à l'intégrateur cross-modal peut améliorer les performances. Par exemple, des références subtiles, du sarcasme ou des contenus socialement nuancés peuvent signaler du sexisme d'une manière qui n'est pas directement codée dans les caractéristiques du texte ou de l'image. Sans ces connaissances externes, le modèle ne peut pas distinguer de manière fiable les contenus sexistes et non sexistes dans les cas limites.

Pour traiter l'inférence dans les cas limites, nous avons intégré des prédictions supplémentaires provenant d'un VLM qui dispose d'une connaissance contextuelle et mondiale plus large, sur les sous-ensembles *S* de l'ensemble de données de validation pour lesquels le modèle renvoie un niveau d'accord donné, défini comme suit :

$$\mathcal{S}_\alpha = \{x \in \text{ValidationSet} \mid \max(\hat{y}_{\text{YES}}(x), \hat{y}_{\text{NO}}(x)) \geq \alpha\} \tag{5}$$

où α indique un seuil de probabilité. Bien que les modèles de Vision-Langage fonctionnent bien dans le cadre d'une évaluation stricte (Nowakowski *et al.*, 2025), ils ne sont pas naturellement adaptés à la modélisation des probabilités de classe. Par conséquent, nous appliquons une prédiction binaire OUI/NON, puis la convertissons en une sortie vectorielle telle que définie ci-dessous :

$$\hat{y}_{VLM} = \begin{cases} (\pi,\ 1-\pi) & \text{si la réponse est OUI} \\ (1-\pi,\ \pi) & \text{autrement} \end{cases} \tag{6}$$

où π est une valeur comprise entre [0, 1]. Ensuite, les prédictions du VLM sont combinées avec les résultats probabilistes du premier modèle ensembliste à l'aide du même produit élémentaire suivi de la stratégie de normalisation décrite dans la Section 6. Cela permet à la prédiction finale de conserver la prédiction du premier modèle ensembliste tout en bénéficiant de la compréhension contextuelle plus riche du VLM sur les mêmes limites ou ambigus.

Le schéma de prédiction qui intègre les résultats des deux modèles est illustré à la Figure 3. Le diagramme fournit un aperçu détaillé du processus décisionnel, en soulignant comment les prédictions générées par l'architecture principale sont d'abord filtrées en termes de probabilités prédites, puis combinées avec les résultats d'un VLM uniquement sur les données limites. La valeur optimale de α, utilisée pour filtrer les prédictions initiales, est étudiée dans la Section 7, dans le but de définir une valeur qui reflète un compromis entre le nombre de mèmes qui doivent être révisés et transmis au VLM et les performances obtenues sur l'ensemble de validation.

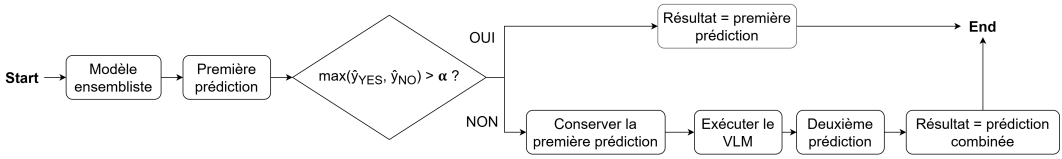

FIGURE 3 – Pipeline décisionnel complet pour la prédiction et la combinaison de labels souples

7 Résultats

En s'appuyant sur l'architecture décrite dans la Section 6, nous avons mené des expériences en utilisant CLIP (Radford *et al.*, 2021) comme encodeur cross-modal. Comme discuté dans la Section 2, des travaux antérieurs ont montré que CLIP capture efficacement le sens sémantique à travers les deux modalités dans les mèmes des réseaux sociaux. Sa principale limitation, cependant, est qu'il a été entraîné principalement sur des données en anglais. Pour y remédier, nous avons également évalué SigLIP 2 (Tschannen *et al.*, 2025), entraîné sur des données multilingues, pour l'encodage multimodal. Nos expériences indiquent néanmoins que CLIP fournit toujours des performances plus cohérentes, probablement en raison des caractéristiques de ses données d'entraînement, mieux adaptées à la nature spécifique de la tâche. Les modèles ont été entraînés sur la prédiction des labels souples, tandis que les labels durs correspondants ont été attribués comme suit :

$$\hat{y}_{\text{Hard}} = \begin{cases} \text{OUI} & \text{si } \hat{y}_{\text{Soft,YES}} > \hat{y}_{\text{Soft,NO}} \\ \text{NON} & \text{autrement} \end{cases} \tag{7}$$

Afin d'étudier la contribution des prédictions binaires produites par le modèle VLM, plusieurs valeurs du paramètre de pondération π, appartenant à l'ensemble :

$$S_\pi = \{0.6, 0.7, 0.8, 0.9, 1.0\}$$

ont été évalués sur l'ensemble de validation. Dans toutes les expériences, le VLM adopté est Qwen 3 VL 32B (Bai *et al.*, 2025), qui est utilisé comme modèle auxiliaire pour affiner les prédictions sous-performantes. Nous avons choisi Qwen VL sur la base de travaux antérieurs réalisés dans le cadre d'EXIST 2025 (Nowakowski *et al.*, 2025), car il combine efficacement la compréhension du texte et de l'image, capture les indices sexistes multimodaux subtils et généralise bien dans les configurations zero-shot et fine-tuned, fournissant des prédictions précises.

Pour plus de détails, le tableau présenté en Annexe A montre les résultats finaux sur l'ensemble de validation pour différentes valeurs de α. On observe que les gains de performance augmentent jusqu'à $\alpha = 0,70$, puis diminuent progressivement, ce qui en fait la valeur de coupure idéale pour déterminer quelles prédictions sont transmises à la supervision par étiquette dure. En complément, en ce qui concerne le poids des prédictions binaires par rapport à la prédiction finale, la configuration la plus performante a été trouvée pour $\pi =$ 0,8. Les résultats sur l'ensemble de test, en adoptant les valeurs $\{\pi, \alpha\}$ mentionnées précédemment, sont présentés dans le Tableau 3.

Les expériences démontrent que le VLM n'est appelé que sur environ 35 % des échantillons testés, tandis que les données restantes sont traitées exclusivement par l'architecture ensembliste sans nécessiter d'appels VLM supplémentaires. En limitant les requêtes VLM aux cas limites uniquement, le cadre proposé permet d'atteindre un compromis favorable entre l'efficacité computationnelle et les performances prédictives, réduisant ainsi considérablement le temps d'inférence. Dans l'ensemble, l'approche obtient des résultats de pointe dans les paradigmes d'évaluation stricte et souples, démontrant des performances cohérentes entre les deux critères d'évaluation. De plus, comme le montre le Tableau 4, l'écart de performance entre les cas anglais et espagnols est presque négligeable, ne représentant en moyenne que Δ = 2,08 % sur les mesures normalisées, ce qui prouve la cohérence de l'approche proposée également sur les mèmes non anglais. Pour plus de détails, la figure présentée en Annexe B fournit un exemple illustrant l'efficacité de l'approche de supervision des étiquettes strictes proposée.

Run	Soft evaluation			Hard evaluation		
	ICM_{Soft}	**ICM_{Soft} Norm**	**Cross Entropy**	**ICM**	**ICM Norm**	**F1**
I2C-Huelva_3 (Carrillo-Casado *et al.*, 2024)	-0.3263	0.4476	1.5189	-0.2772	0.4036	0.4714
VictorUNED_1 (Ruiz *et al.*, 2024)	-0.2925	0.4530	1.1028	0.0641	0.5326	0.7051
TrankilTwice_3 (Italiani *et al.*, 2025)	-0.2198	0.4652	1.0394	0.0562	0.5303	0.6942
Modèle ensembliste	-0.1832	0.4711	**0.9615**	0.2331	0.6193	0.7492
+ supervision du VLM	**-0.0583**	**0.4908**	0.9858	**0.3216**	**0.6668**	**0.7850**

TABLE 3 – Résultats sur l'ensemble de test après application de la supervision VLM aux mèmes limites (performance moyenne sur 3 seeds)

Langue	Soft evaluation			Hard evaluation		
	ICM_{Soft}	ICM_{Soft} Norm	Cross Entropy	ICM	ICM Norm	F1(YES)
Anglais	**-0.0416**	**0.4934**	0.9906	**0.3279**	**0.6664**	0.8024
Espagnol	-0.1231	0.4808	**0.9815**	0.2907	0.6556	**0.8473**

TABLE 4 – Résultats finaux sur l'ensemble de test, regroupés par langue (performance moyenne sur 3 seeds)

8 Conclusion et perspectives

L'analyse expérimentale démontre que le cadre proposé combine efficacement l'adaptabilité des architectures basées sur les transformateurs pour la prédiction des étiquettes souples avec les connaissances contextuelles des modèles visuels-linguistiques, établissant ainsi une nouvelle référence sur les benchmarks évalués. Les VLM sont naturellement enclins à produire des étiquettes strictes, ce qui correspond à la décision binaire requise par la tâche EXIST, mais limite la capture de l'incertitude des cas sexistes nuancés. Les architectures basées sur des transformateurs, en revanche, produisent des résultats probabilistes qui reflètent l'ambiguïté dans l'étiquetage, explicitement valorisée par le protocole d'évaluation souple d'EXIST.

L'**efficacité et la nouveauté du cadre proposé** reposent sur deux aspects principaux :

— l'introduction d'une **stratégie d'entraînement à double partition** dans le but de prédire des étiquettes souples, en utilisant des ensembles déséquilibrés. L'approche démontre que l'acquisition de connaissances à partir de partitions déséquilibrées peut conduire à de meilleures performances lors de l'intégration des sorties d'un modèle dans une stratégie d'ensemble, sans recourir à des connaissances externes (par exemple, la génération de descriptions de mèmes ou la recherche agentique sur les images)

— l'**intégration mutuelle de prédictions probabilistes et déterministes** au moyen d'un mécanisme de routage sensible aux probabilités, qui déclenche sélectivement un VLM uniquement pour les mèmes associés à une forte incertitude prédictive. Ce mécanisme permet au système de fournir des probabilités pondérées bien calibrées sans compromettre l'efficacité computationnelle

L'approche proposée satisfait aux deux exigences d'évaluation de la Tâche 2.1 d'EXIST 2025 : elle maintient de solides performances dans le cadre d'une évaluation à étiquettes strictes tout en améliorant simultanément les mesures à étiquettes souples en modélisant explicitement l'accord entre les annotateurs, contribuant ainsi à la durabilité sociale en favorisant un traitement équitable et inclusif des divers points de vue. Cette stratégie de prédiction unifiée démontre que la méthode proposée concilie avec succès la classification déterministe et le raisonnement probabiliste, remédiant ainsi à une limitation majeure des approches existantes dans le cadre de la tâche. Sur la base de l'approche proposée, les travaux futurs pourraient explorer des critères de partitionnement plus avancés qui utilisent à la fois les distributions de classes et les caractéristiques spécifiques aux mèmes (texte et visuels). En outre, l'extension de la supervision de niveau dur pour influencer les états cachés du modèle pourrait combiner les prédictions souples et dures dans les intégrations, améliorant ainsi la capture des relations complexes et renforçant les performances prédictives. Ces orientations pourraient améliorer l'intégration des informations multimodales et des stratégies de supervision pour détecter le sexisme dans les contenus des médias sociaux.

Références

AMIGO E. & DELGADO A. (2022). Evaluating extreme hierarchical multi-label classification. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 5809–5819, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.399.

BAI S., CAI Y., CHEN R., CHEN K., CHEN X., CHENG Z., DENG L., DING W., GAO C., GE C., GE W., GUO Z., HUANG Q., HUANG J., HUANG F., HUI B., JIANG S., LI Z., LI M., LI M., LI K., LIN Z., LIN J., LIU X., LIU J., LIU C., LIU Y., LIU D., LIU S., LU D., LUO R., LV C., MEN R., MENG L., REN X., REN X., SONG S., SUN Y., TANG J., TU J., WAN J., WANG P., WANG P., WANG Q., WANG Y., XIE T., XU Y., XU H., XU J., YANG Z., YANG M., YANG J., YANG A., YU B., ZHANG F., ZHANG H., ZHANG X., ZHENG B., ZHONG H., ZHOU J., ZHOU F., ZHOU J., ZHU Y. & ZHU K. (2025). Qwen3-vl technical report.

CARRILLO-CASADO , ROMÁN-PÁSARO J., MATA-VÁZQUEZ J. & PACHÓN-ÁLVAREZ V. (2024). I2c-uhu at exist 2024 : Transformer-based detection of sexism and source intention in memes using a learning with disagreement approach. In G. FAGGIOLI, N. FERRO, P. GALUŠČÁKOVÁ & A. G. S. DE HERRERA, Éds., *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, CEUR-WS.org, p. 978–992 : CEUR Workshop Proceedings. CLEF 2024, Grenoble, France.

GOODFELLOW I., BENGIO Y. & COURVILLE A. (2016). *Deep Learning*. Cambridge, MA : MIT Press. Chapter 5 discusses cross-entropy loss.

HAKIMOV S., CHEEMA G. S. & EWERTH R. (2022). Tib-va at semeval-2022 task 5 : A multimodal architecture for the detection and classification of misogynous memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, p. 756–760.

ITALIANI P., MAQBOOL F., GIMENO-GÓMEZ D., FERSINI E. & MARTÍNEZ-HINAREJOS C.-D. (2025). Trankiltwice at exist2025 : Detecting sexism in memes under multi-lingual settings. In G. FAGGIOLI, N. FERRO, P. GALUŠČÁKOVÁ & A. G. S. DE HERRERA, Éds., *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum*, CEUR-WS.org, p. 2012–2022 : CEUR Workshop Proceedings. Notebook for the EXIST Lab at CLEF 2025, Madrid, Spain.

MOSTAFAZADEH DAVANI A., DÍAZ M. & PRABHAKARAN V. (2022). Dealing with disagreements : Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, **10**, 92–110. DOI : 10.1162/tacl_a_00449.

NAEBZADEH A., NOBAKHTIAN M. & EETEMADI S. (2024). Nica at exist clef tasks 2024. In *Conference and Labs of the Evaluation Forum*.

NOWAKOWSKI N., CALOGIURI L., EGYED-ZSIGMOND E., NURBAKOVA D., ERBANI J. & CALABRETTO S. (2025). GrootWatch at EXIST 2025 : Automatic Sexism Detection on Social Networks - Classification of Tweets and Memes. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, Madrid, Spain. HAL : hal-05157015.

PAN R., BERNAL BELTRÁN T., GARCÍA DÍAZ J. A. & VALENCIA-GARCÍA R. (2025). Umuteam at exist 2025 : multimodal transformer architectures and soft-label learning for sexism detection. *Working Notes of CLEF*.

PLAZA L., CARRILLO-DE ALBORNOZ J., ARCOS I., ROSSO P., SPINA D., AMIGÓ E., GONZALO J. & MORANTE R. (2025). Overview of exist 2025 : Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction : 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9–12, 2025, Proceedings*, p. 266–289, Berlin, Heidelberg : Springer-Verlag. DOI : 10.1007/978-3-032-04354-2_16.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G. & SUTSKEVER I. (2021). Learning transferable visual models from natural language supervision.

RUIZ V., DE ALBORNOZ J. C. & PLAZA L. (2024). Concatenated transformer models based on levels of agreements for sexism detection. In G. FAGGIOLI, N. FERRO, P. GALUŠČÁKOVÁ & A. G. S. DE HERRERA, Éds., *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, CEUR-WS.org, p. 1187–1197 : CEUR Workshop Proceedings. Notebook for the EXIST Lab at CLEF 2024, Grenoble, France.

TSCHANNEN M., GRITSENKO A., WANG X., NAEEM M. F., ALABDULMOHSIN I., PARTHASARATHY N., EVANS T., BEYER L., XIA Y., MUSTAFA B., HÉNAFF O., HARMSEN J., STEINER A. & ZHAI X. (2025). Siglip 2 : Multilingual vision-language encoders with improved semantic understanding, localization, and dense features.

ZHAO H., YUAN C., HUANG F., HU X., ZHANG Y., YANG A., YU B., LIU D., ZHOU J., LIN J., YANG B., CHENG C., TANG J., JIANG J., ZHANG J., XU J., YAN M., SUN M., ZHANG P., XIE P., TANG Q., ZHU Q., ZHANG R., WU S., ZHANG S., HE T., TANG T., XIA T., LIAO W., SHEN W., YIN W., ZHOU W., YU W., WANG X., DENG X., XU X., ZHANG X., LIU Y., LI Y., ZHANG Y., JIANG Y., WAN Y. & ZHOU Y. (2025). Qwen3guard technical report.

Annexe A Résultats finaux obtenus sur l'ensemble de validation avec différentes valeurs de α

	Soft evaluation		Hard evaluation	
α	**ICM_{Soft} Norm**	**Cross Entropy**	**ICM Norm**	**F1(YES)**
0.60	0.4807 ± 0.02	0.9796 ± 0.05	0.6276 ± 0.04	0.8034 ± 0.04
0.65	0.4940 ± 0.02	0.9833 ± 0.05	0.6630 ± 0.05	0.8210 ± 0.04
0.70	0.4957 ± 0.02	0.9975 ± 0.03	0.6621 ± 0.03	0.8189 ± 0.02
0.75	0.5004 ± 0.02	1.0065 ± 0.03	0.6677 ± 0.03	0.8082 ± 0.05
0.80	0.5043 ± 0.02	1.0201 ± 0.03	0.6655 ± 0.03	0.8160 ± 0.03

TABLE A1 – Résultats obtenus sur l'ensemble de validation en adoptant la révision VLM avec différentes valeurs de α (moyenne ± écart-type sur 3 graines). Les résultats montrent que l'adoption d'une valeur de α = 0,70 permet d'obtenir un compromis favorable entre les performances et le nombre de prédictions à réviser (environ 40 % par rapport à la taille de l'ensemble de validation).

Annexe B Exemple du système de révision basé sur le VLM

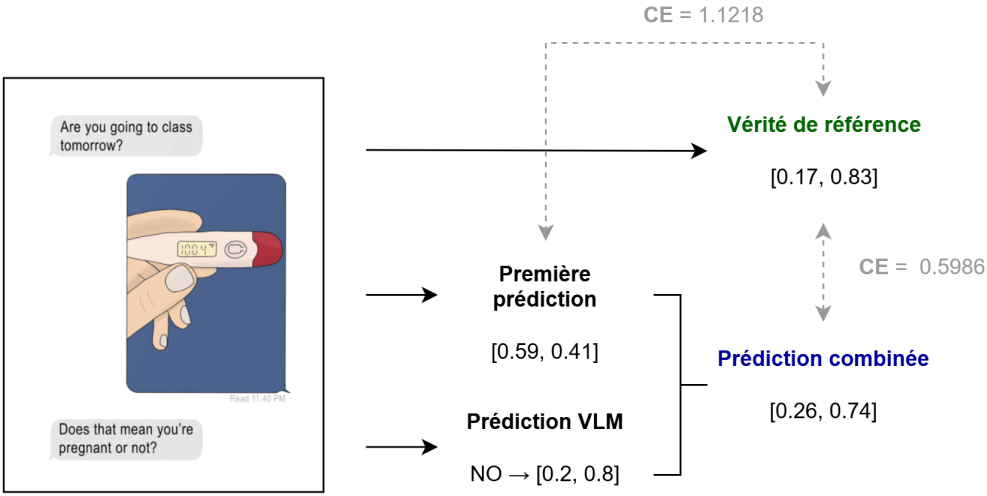

FIGURE 4 – Exemple du système de révision basé sur un VLM avec des étiquettes strictes pour un mème limite, caractérisé par une faible différence entre les probabilités prédites de OUI et NON (avec $\alpha = 0.70$ et $\pi = 0.8$). Dans la figure, CE désigne l'entropie croisée calculée entre la prédiction et la vérité terrain. À partir d'une prédiction initiale orientée vers une attribution d'étiquette incorrecte, la réponse VLM est d'abord vectorisée, puis combinée avec la sortie probabiliste d'origine. La fusion des deux sorties conduit à une prédiction finale nettement plus proche de la vérité terrain du mème présenté, pour lequel seul un annotateur sur six a marqué le contenu comme sexiste.