Au-delà du Texte : Intégrer le Contexte Web pour Modérer le Sexisme en Ligne

Nathan Nowakowski[1] Elöd Egyed-Zsigmond[1] Sylvie Calabretto[1] Diana Nurbakova[1]
(1) INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69621 Villeurbanne, France
{nathan.nowakowski, elod.egyed-zsigmond, sylvie.calabretto, diana.nurbakova}@insa-lyon.fr

RÉSUMÉ

Le sexisme sur les réseaux sociaux représente un défi croissant pour la modération automatique, notamment en raison de sa nature implicite et de son évolution constante. Cet article présente une approche novatrice qui combine la recherche d'information en temps réel sur Internet et un apprentissage multitâche pour capturer les nuances fines du sexisme. En enrichissant les modèles BERT avec un contexte externe dynamique, notre méthode permet de lever les ambiguïtés liées à l'actualité. Soumise au challenge international EXIST 2025, notre solution s'est classée première pour l'évaluation en soft labels. Ces résultats démontrent que l'intégration du contexte situationnel est essentielle pour modérer efficacement la complexité des échanges sur les médias sociaux.

Avertissement : Ce document contient des exemples de propos haineux, explicites et sexistes présentés à des fins illustratives.

ABSTRACT

Beyond Text : Integrating Web Context to Moderate Online Sexism

Sexism on social media represents a growing challenge for automated moderation, particularly due to its implicit nature and constant evolution. This paper introduces an innovative approach that combines real-time Internet information retrieval with multi-faceted learning to capture the fine-grained nuances of sexism. By enriching BERT models with dynamic external context, our method effectively resolves ambiguities linked to current events and cultural trends. Submitted to the EXIST 2025 international challenge, our solution ranked first in the soft labels evaluation. These results demonstrate that integrating situational context is essential for effectively moderating the complexity of social media discourse.

Content Warning : This paper includes examples of hateful, explicit and sexist language presented for illustrative purposes.

MOTS-CLÉS : Détection de Sexisme, Classification de Texte, Traitement Automatique du Language, Transformers.

KEYWORDS: Sexism Detection, Text Classification, Natural Language Processing, Transformers.

ARTICLE ACCEPTÉ À : Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025).
URL : https://ceur-ws.org/Vol-4038/paper_160.pdf

1 Introduction

Le sexisme, sous forme de préjugés ou de commentaires haineux, est une forme répandue de violence numérique qui doit être combattue dans un contexte où les réseaux sociaux et les plateformes numériques sont omniprésents. En 2026, les femmes représentent 84% des victimes de cybersexisme sur les réseaux sociaux en France (Haut Conseil à l'Egalité entre les Femmes et les Hommes, 2026). Cette situation préoccupante représente un défi sociétal majeur, qui consiste à trouver un équilibre entre les attentes éthiques en matière de modération et la nécessité de protéger la liberté d'expression. Ce travail s'inscrit dans un contexte où des plateformes telles que Meta assouplissent considérablement leurs politiques de modération, exacerbant les risques de polarisation et de haine sexiste (Meta, 2025; Amnesty International, 2025). Dans le même temps, le discours masculiniste gagne en visibilité, rendant indispensable le développement d'outils capables de cartographier et de contrer ces dynamiques en temps réel (Institut du Genre en Géopolitique and Equipop, 2023).

Par conséquent, l'identification automatique des contenus sexistes sur les réseaux sociaux devient une tâche cruciale. Afin d'encourager de telles initiatives, le défi EXIST 2025 (Plaza *et al.*, 2025a,b) comprend neuf sous-tâches en deux langues, l'anglais et l'espagnol, qui sont les trois mêmes tâches appliquées à trois types de données différents : texte (tweets), image (memes) et vidéo (TikToks). Dans le cadre de ce travail, nous nous concentrons exclusivement sur la **modalité textuelle**. Les définitions du sexisme adoptés suivent la taxonomie officielle de la campagne EXIST 2025 :

1. **Niveau 1 : Détection Binaire (Identification)** : Cette tâche vise à discriminer les contenus sexistes des contenus neutres. Elle établit la présence d'un biais de genre, d'une hostilité ou d'une discrimination (**Sexiste** vs **Non-Sexiste**).
2. **Niveau 2 : Analyse de la Source et de l'Intention** : Pour les messages identifiés comme sexistes, ce niveau caractérise la posture discursive de l'auteur. On distingue le sexisme **Direct** (agression immédiate), le sexisme **Rapporté** (reprise de propos tiers, souvent à des fins de dénonciation ou de partage) et le sexisme de **Jugement** (jugements de valeur sur les comportements ou l'apparence).
3. **Niveau 3 : Caractérisation Granulaire** : Ce niveau final consiste en une classification multi-label permettant de capturer la nature complexe de l'offense, pour les tweets identifiés comme sexistes. Les catégories couvrent le spectre de l'inégalité : **Idéologie et Inégalité** (II), **Stéréotypes et Domination** (SD), **Objectivation** (OBJ), **Violence Sexuelle** (VS) et **Misogynie et Violence non-sexuelle** (MVNS).

Pour modéliser efficacement ces différentes catégories au sein d'un unique framework, nous proposons une architecture multitâche unifiée qui utilise des informations contextuelles externes pour améliorer la compréhension sémantique des tweets.

2 Etat de l'Art

L'évolution de la détection du sexisme en ligne reflète une transition paradigmatique, passant de l'analyse lexicale de surface à la modélisation profonde des nuances contextuelles. Historiquement, les travaux se sont appuyés sur des approches classiques d'apprentissage automatique, telles que les SVM ou les forêts aléatoires, dépendantes d'une ingénierie de caractéristiques manuelles (*N-grams*, TF-IDF) (Chhabra & Vishwakarma, 2023). Bien que robustes pour identifier des termes

explicitement haineux, ces méthodes se heurtaient à la plasticité du langage naturel et à l'évolution constante des codes linguistiques sur les réseaux sociaux. Pour pallier ce manque de flexibilité, l'introduction des architectures neuronales a permis de capturer des structures sémantiques plus fines. Les modèles hybrides, notamment les architectures CNN-BiLSTM, ont marqué une étape clé en combinant l'extraction de motifs locaux et la capture de dépendances séquentielles à long terme (Vetagiri *et al.*, 2025). Toutefois, la véritable rupture conceptuelle survient avec le mécanisme d'attention (Vaswani *et al.*, 2017) qui a permis de contextualiser chaque unité lexicale au sein de l'intégralité du message.

Cette ère des modèles pré-entraînés, initiée par BERT (Devlin *et al.*, 2019), a permis d'affiner la détection des sous-entendus sexistes grâce à des variantes optimisées comme RoBERTa (Liu *et al.*, 2019) ou DeBERTa (He *et al.*, 2021; Fang *et al.*, 2024). Parallèlement, l'émergence des modèles de langage de grande taille (LLM) tels que Llama-3 (Grattafiori *et al.*, 2024), optimisables via des méthodes d'adaptation de bas rang (LoRA) (Hu *et al.*, 2022), ouvre des perspectives inédites sur le raisonnement contextuel et l'interprétation du sarcasme (Quan & Thin, 2024). Enfin, des recherches récentes soulignent que la détection du sexisme ne peut s'extraire de la dimension affective du discours ; l'intégration de l'analyse des sentiments permet ainsi de lever l'ambiguïté sur des commentaires sexistes dissimulés sous une apparente neutralité émotionnelle (Belbachir *et al.*, 2024).

Évolution des Ressources et Paradigmes d'Annotation La qualité de la détection automatique dépend intrinsèquement de la richesse sémantique et de la fiabilité sociologique des données d'entraînement. L'évolution des jeux de données illustre une transition vers une prise en compte accrue de la subjectivité et de la complexité du sexisme.

Initialement, les corpus comme celui de Debnath *et al.* (2020) se concentraient sur la classification thématique à grande échelle (ex : #metoo). La nécessité d'une granularité accrue a ensuite mené à des taxonomies hiérarchisées, comme celle de SemEval-2023 Task 10 (Kirk *et al.*, 2023), introduisant des catégories fines (stéréotypes, harcèlement) pour une détection plus explicable.

Parallèlement, la recherche actuelle remet en question l'existence d'une vérité terrain unique. Les campagnes EXIST (2021-2025) (Rodríguez-Sánchez *et al.*, 2021, 2022; Plaza *et al.*, 2023, 2024, 2025a) sont pionnières dans l'adoption du *perspectivisme*. En fournissant des données démographiques sur les annotateurs et en intégrant le paradigme *Learning with Disagreements* (LeWiDi) (Leonardelli *et al.*, 2023), EXIST permet de modéliser le sexisme comme un phénomène dont la perception varie selon le genre, l'âge et la culture de l'observateur. Cette approche transforme le désaccord entre annotateurs, autrefois considéré comme un bruit, en un signal riche pour la robustesse des modèles.

Limites des Approches Actuelles et Positionnement Malgré les avancées significatives permises par les architectures transformatrices, la détection automatique du sexisme se heurte, entre autres, à deux verrous scientifiques que cette étude entend lever :

1. **Désalignement temporel et linguistique :** Les modèles de langue pré-entraînés souffrent d'une rigidité sémantique face à la vélocité des réseaux sociaux. Comme le souligne Valavi *et al.* (2022), l'évolution des codes langagiers et l'émergence de nouveaux néologismes sexistes, souvent utilisés pour contourner la modération, rendent les poids des modèles rapidement obsolètes. Ce phénomène de glissement temporel limite l'efficacité des classifieurs statiques face aux tendances émergentes sur des plateformes comme X.
2. **Pluralité taxonomique et subjectivité de l'observateur :** Le sexisme n'est pas un phénomène binaire, mais un spectre complexe incluant le harcèlement, la misogynie, ou encore le sexisme narratif et discursif. Par ailleurs, comme le souligne le paradigme LeWiDi (Leonardelli

et al., 2023), la perception d'un contenu est intrinsèquement liée à l'identité de l'observateur. L'absence de prise en compte des métadonnées des annotateurs et de la hiérarchie des offenses conduit souvent les modèles à imposer un consensus artificiel qui masque la réalité sociologique du phénomène (Plaza *et al.*, 2024).

Dans cet article, nous proposons une approche multidimensionnelle pour répondre à ces défis. Notre méthodologie, présentée en section 4, introduit trois contributions majeures : (i) un mécanisme d'enrichissement par **recherche d'information en temps réel**, permettant de prendre com compte le contexte externe du tweet à sa classification ; (ii) un cadre d'**apprentissage hiérarchique** conçu pour modéliser la granularité des types de sexisme ; (iii) l'intégration systématique des **métadonnées des annotateurs** au sein du processus de classification, afin de capturer et de respecter la diversité des perceptions humaines face au sexisme.

3 Données et Protocole Expérimental

Cette étude exploite donc le corpus de la campagne d'évaluation EXIST 2025. La Table 1 synthétise la distribution des données. Notre protocole suit la structure suivante : le *fine-tuning* est opéré sur l'ensemble d'entraînement, tandis que la validation interne permet d'évaluer nos itérations méthodologiques. Les performances finales sont mesurées sur l'ensemble de test officiel, dont les étiquettes réelles (*gold labels*) demeurent masquées, garantissant l'intégrité de la comparaison avec les autres participants du challenge.

TABLE 1 – EXIST 2025 Tweets : Répartition linguistique et partitionnement (en nombre de tweets)

Langage	Entraînement	Evaluation	Test
Anglais	3260	489	978
Espagnol	3660	549	1098

Métriques d'Évaluation et Philosophie de l'ICM La performance est quantifiée par la mesure de contraste de l'information (*Information Contrast Measure*, ICM) (Amigo & Delgado, 2022), et plus spécifiquement sa variante normalisée, **ICM-Norm**. Contrairement au F1-score qui pénalise de manière uniforme toute divergence par rapport à une vérité binaire, l'ICM intègre une dimension probabiliste essentielle dans un contexte hiérarchique. Dans notre approche, une erreur de détection au Niveau 1 (omission d'un contenu sexiste) est pénalisée plus lourdement qu'une confusion de catégorie au Niveau 3 (ex : méprise entre *Misogynie* et *Stéréotypes*). L'ICM permet ainsi de valoriser les modèles qui, faute de certitude, capturent correctement la hiérarchie sexiste, plutôt que ceux qui imposent une classification arbitraire.

Spécificités du Jeu de Données et Gestion de l'Incertitude Une particularité notable du corpus EXIST 2025 réside dans son protocole d'annotation : chaque entrée est évaluée par un panel de six annotateurs. Cette configuration paire autorise des situations de parité stricte (ex. : 3 *Oui* contre 3 *Non*), rendant l'inférence d'une étiquette unique (*hard label*) scientifiquement contestable. Dans la Table 2, ces instances sont explicitement marquées comme « Non étiquetées », signalant une indécidabilité sémantique ou une divergence de perception irréductible. Plutôt que d'écarter ces données, notre approche les intègre comme un signal d'apprentissage à part entière ; les modalités techniques de leur traitement algorithmique sont détaillées dans la Section 4.

TABLE 2 – Distribution des Classes des Données d'Entraînement (en pourcentage %)

Niveau 1			**Non**	**Oui**	**Non étiquetées**		
			48.65	38.97	12.37		
Niveau 2		**NON**	**Direct**	**Jugement**	**Rapporté**	**Non étiquetées**	
		48.65	18.70	5.43	6.63	20.58	
Niveau 3	**Non**	**II**	**SD**	**OBJ**	**SV**	**MNSV**	**Non étiquetées**
	48.65	16.08	20.56	20.56	22.12	12.37	12.57

4 Méthodologie

4.1 Représentation des Données et Profilage des Annotateurs

La nature subjective du sexisme impose une représentation qui dépasse le simple contenu textuel. Notre approche articule le nettoyage lexical et l'intégration de variables sociodémographiques. Pour les modèles de type *Encoder* (BERT), le texte brut des tweets subit une normalisation spécifique aux réseaux sociaux afin de limiter le bruit sans altérer la charge sémantique : suppression des mentions utilisateurs, des URLs et des emojis, ces derniers n'apportant pas d'amélioration significative selon nos expériences. Pour les modèles génératifs (LLM), nous conservons une structure textuelle quasi-brute, incluant les emojis, conformément aux préconisations de Quan & Thin (2024), afin de préserver le signal pragmatique. La Table 3 synthétise les étapes appliquées.

TABLE 3 – Processus de nettoyage des données pour le modèle BERT : suppression des mentions, des URL, des espaces superflus, des emojis et des caractères HTML, conversion du texte en minuscules et décensure des mots (identification et correction des mots censurés, par exemple « *f**k* » → « *fuck* »).

Avant formatage	Arpès formatage
Feel #blessed that I have raised a caring & loving 13 yo who is our Next Gen Feminist & Ally. I was crying inside when I got this text. Not only we must #BreakTheBias for women, we need to do it for our children. 🗼🍣💎🧋 @GlobalFundWomen @UN_Women @womensday @WomeninID https://t.co/UJvloR0IP	feel blessed that i have raised a caring loving 13 yo who is our next gen feminist ally. i was crying inside when i got this text. not only we must breakthebias for women, we need to do it for our children.

Conformément au paradigme *perspectiviste*, nous ne cherchons pas à effacer le désaccord entre annotateurs, mais à le modéliser. Une analyse statistique préalable (tests du χ^2 et régressions logistiques) a permis d'identifier les variables ayant l'impact le plus significatif sur la perception du sexisme : le **niveau d'études**, le **pays d'origine** et l'**origine ethnique**. Pour chaque tweet, les profils des annotateurs sont agrégés et transformés en un vecteur de caractéristiques de dimension $d = 65$. Cette dimension résulte de la concaténation des encodages one-hot des trois variables retenues : 7 catégories d'ethnicité, 6 niveaux d'études et 52 pays représentés dans le corpus. Ce vecteur, illustré par un exemple fictif dans la Figure 1, est concaténé à la représentation latente issue du jeton `[CLS]` de BERT avant d'être transmis aux couches de classification. Cette méthode permet au modèle d'ajuster

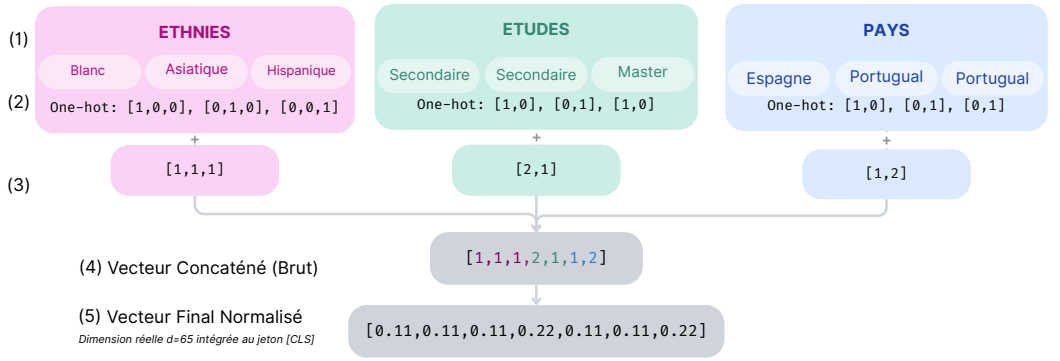

FIGURE 1 – **Vectorisation des Métadonnées Annotateurs** (exemple fictif). (1) Chaque tweet évalué comprend des informations d'annotation classées par catégorie. (2) Chaque catégorie est codée en one-hot. (3) Pour chaque catégorie, les vecteurs codés en one-hot correspondants sont additionnés pour générer un vecteur de catégorie. (4) Les vecteurs de catégorie créés sont concaténés en un seul vecteur d'informations sur l'annotateur. (5) Enfin, le vecteur résultant est normalisé.

sa prédiction en fonction du contexte sociodémographique du panel d'annotation, capturant ainsi la variabilité inhérente aux jugements humains.

4.2 Motivation : Analyse des Limites Statiques

Une évaluation préliminaire de nos modèles de base (BERT et LLM) a révélé un plafond de performance lié à la *cécité contextuelle*. L'analyse de l'espace latent via TF-IDF a mis en évidence que certains termes pivots (ex : « femme », « enceinte ») dominent indûment les prédictions, conduisant à des faux positifs systématiques sur des contenus pourtant neutres mais faisant référence à des événements culturels récents (ex : mèmes viraux). Pour lever ce verrou, nous introduisons une architecture de Recherche Augmentée par Génération (RAG) orchestrée par un agent IA autonome. Notre agent intervient de manière sélective selon le flux de travail suivant :

1. **Évaluation de la complétude :** L'agent analyse le tweet pour déterminer si une connaissance extrinsèque est requise pour son interprétation.
2. **Récupération dynamique :** En cas d'ambiguïté, l'agent formule des requêtes optimisées vers le moteur de recherche DuckDuckGo pour capturer le contexte immédiat (actualités, tendances, références argotiques).
3. **Synthèse contextuelle :** Les informations brutes sont distillées par le LLM de l'agent pour n'en extraire que les éléments sémantiques pertinents.

Nous utilisons le modèle `Llama-3.3-70B-Instruct` pour piloter cet agent en raison de ses capacités supérieures de planification et de manipulation d'outils (Wang *et al.*, 2024). La Figure 2 détaille le fonctionnement de cet agent IA, et le prompt utilisé est présenté dans l'Annexe A. Nous avons donc, pour chaque tweet du dataset EXIST 2025, fait appel à cet agent IA afin de créer une base de données de leur contexte.

L'intégration du contexte récupéré à la détection automatique du sexisme a été testée selon trois

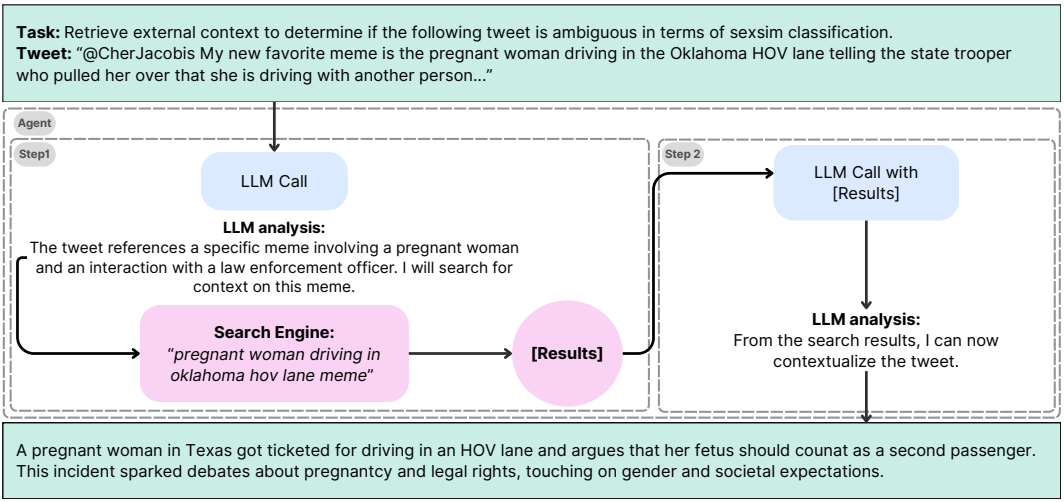

FIGURE 2 – Fonctionnement de l'agent IA analysant les références d'un tweet à l'aide d'un appel externe vers un moteur de recherche.

paradigmes :

- **Architecture Siamese Dual Encoder (SDE) :** Pour les modèles BERT, nous avons implémenté une structure à deux encodeurs (Dong *et al.*, 2022) permettant de projeter indépendamment le tweet et son contexte dans un espace sémantique commun avant leur fusion.
- **Inférence directe par l'Agent :** Le modèle `Llama-3.3-70B` effectue une classification *few-shot* en intégrant directement les résultats de recherche dans son raisonnement interne. Voir l'Annexe B pour le prompt utilisé.
- **Fine-tuning avec contexte :** Le modèle `Llama-3.2-3B` est affiné sur le corpus EXIST avec l'injection du contexte directement dans le *prompt* d'instruction. Voir Annexe C pour le prompt utilisé.

La Table 4 présente les résultat préliminaires sur les données d'évaluation. Les configurations spécifiques utilisées sont détaillées dans l'Annexe D.

TABLE 4 – Comparaison des différentes approches pour la tâche 1 sur les données d'évaluation.

Modèle	Configuration	ICM-Hard Norm
Llama-3.3-70B-Instruct	Agent IA seul	0,69
Llama-3.2-3B-Instruct	Fine-tuné + Contexte	0,60
XLM Roberta Large	Fine-tuné + Contexte + Informations Annotateurs	**0,81**

Les expériences révèlent que l'architecture SDE basée sur BERT offre des performances supérieures aux approches basées sur les LLMs. Nous attribuons la sous-performance des LLMs au phénomène de détournement de contexte (*context hijacking*) (Jeong, 2023). Dans ce scénario, le modèle de langage accorde une importance disproportionnée aux informations externes récupérées, au détriment des nuances linguistiques subtiles présentes dans le tweet original, ce qui dégrade la précision de la classification. En conséquence, la suite de notre méthodologie se concentre sur l'optimisation de

l'architecture basée sur BERT, qui offre un meilleur compromis entre précision lexicale et intégration du contexte dynamique. Par souci de transparence méthodologique, une évaluation préliminaire de la qualité des contextes générés est fournie en Annexe E. Bien que l'analyse approfondie de la fidélité des informations récupérées dépasse le cadre de cette étude, nous reconnaissons qu'elle constitue une perspective d'investigation cruciale pour la robustesse des futurs systèmes de détection contextuelle.

4.3 Architecture BERT multitâche Hiérarchique

Pour modéliser la complexité taxonomique du sexisme, nous proposons une architecture BERT multitâche optimisée, conçue pour traiter simultanément les trois niveaux de classification tout en intégrant des étiquettes dures et souples (Stickland & Murray, 2019). Cette approche favorise le transfert de connaissances entre les tâches via des couches d'encodage partagées, tout en préservant la spécificité sémantique de chaque niveau par des têtes de classification dédiées (Peng *et al.*, 2020).

Inspirés par Fang *et al.* (2024), nous avons fait évoluer la structure classique vers une architecture dite `2+1` (cf. Figure 3). Cette partition est dictée par la divergence des espaces de sortie : **(A) la Tête A (Exclusivité Sémantique)**, dédiée aux tâches 1 (Détection) et 2 (Intention), utilise une fonction *Softmax* pour traiter des catégories mutuellement exclusives ; **(B) la Tête B (Multi-label)**, dédiée à la tâche 3 (Catégorisation), emploie une fonction *Sigmoid* afin de modéliser la co-occurrence possible de plusieurs formes de sexisme (ex. : *Objectivation* et *Stéréotypes*) au sein d'un même tweet.

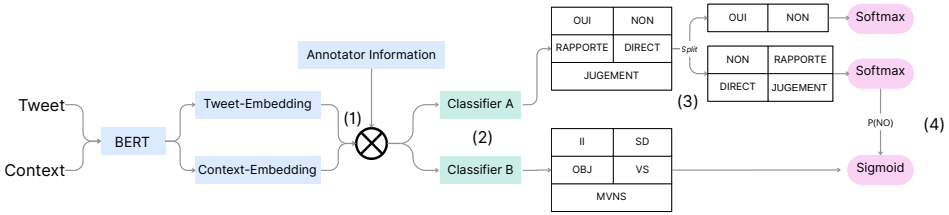

FIGURE 3 – **Architecture Multitâche** : (1) Fusion par concaténation des *embeddings* SDE (tweet/contexte) et du vecteur sociodémographique des annotateurs. (2) Têtes de classification dédiées opérant sur la représentation latente globale. (3) Partitionnement de la Tête A pour les tâches mutuellement exclusives. (4) Injection de la probabilité de la classe « NON » dans la Tête B pour assurer la cohérence hiérarchique de la tâche 3.

Notre innovation architecturale réside aussi dans l'exploitation de la cohérence sémantique de la classe *NON* (non-sexiste) à travers les trois niveaux de classification. Pour matérialiser cette dépendance, l'entraînement de notre architecture `2+1` repose sur une fonction de perte globale. Soit N la taille du lot (*batch size*), la perte totale $\mathcal{L}_{total}$ est définie par :

$$\mathcal{L}_{total} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 \tag{1}$$

Les pertes $\mathcal{L}_1$ (Détection binaire) et $\mathcal{L}_2$ (Intention) sont calculées via une entropie croisée standard sur les logits de la Tête A. Le calcul de $\mathcal{L}_3$ est conditionné par un masquage hiérarchique. Le masque M_i s'active exclusivement pour les instances appartenant à la classe sexiste prédite par la tête A ($\hat{p}_{i,yes} > 0.5$) :

$$M_i = \mathbb{I}(\hat{p}_{i,yes} > 0.5) \tag{2}$$

La perte de la tâche de Niveau 3 (*Multi-label*) est alors calculée via une entropie croisée binaire (BCE) conditionnelle, strictement restreinte aux échantillons classés sexistes :

$$\mathcal{L}_3 = \frac{1}{\sum_{i=1}^{N} M_i + \epsilon} \sum_{i=1}^{N} M_i \cdot \text{BCE}(\hat{y}_i^{(3)}, y_i^{(3)}) \tag{3}$$

Où ϵ est une constante de stabilité numérique (10^{-10}). Ce filtrage permet de concentrer la capacité discriminatoire du modèle sur les nuances de l'offense.

Contrairement aux approches multitâche classiques qui exigent l'ajustement empirique de coefficients de pondération des différentes pertes, notre architecture maintient une pondération unitaire. Ce choix de conception s'aligne intrinsèquement sur la philosophie de la métrique ICM (Amigo & Delgado, 2022). Par nature, l'architecture impose sa propre pondération hiérarchique : une erreur au Niveau 1 (incapacité à détecter le sexisme) invalide structurellement la pertinence des niveaux inférieurs, reflétant la lourde pénalité imposée par l'ICM, par opposition à une simple erreur de granularité (confusion entre deux sous-catégories au Niveau 3).

Apprentissage par Étiquettes Souples (Soft Label Learning) L'un des défis majeurs du corpus EXIST 2025 réside dans la gestion des données « Non étiquetées » lors d'une parité stricte (ex. 3 voix *Oui* contre 3 voix *Non*). Dans un paradigme d'apprentissage par étiquettes strictes (*Hard Label Learning*, HLL), ces instances sont systématiquement écartées, entraînant une perte d'environ 10% du signal d'entraînement pour le Niveau 1, et davantage pour les niveaux granulaires (Sanchez & Zhang, 2022). Pour pallier cette perte d'information, nous adoptons une approche d'apprentissage par étiquettes souples (*Soft Label Learning*, SLL) (de Vries & Thierens, 2024). Au lieu de réduire le vote du panel à une classe unique, nous entraînons le modèle à prédire la distribution de probabilités issue des annotations. Par exemple, un tweet ayant recueilli 5 votes *Oui* et 1 vote *Non* est modélisé par le vecteur cible $[0,83; 0,17]$. Cette méthode permet non seulement d'exploiter l'intégralité du corpus, mais aussi de capturer la nuance et l'incertitude inhérentes au paradigme LeWiDi.

En synthèse, notre méthodologie articule l'enrichissement contextuel par agents IA, l'intégration des métadonnées sociodémographiques et une architecture multitâche hiérarchique. Cette synergie permet de transformer la détection du sexisme d'une simple classification lexicale en une analyse contextuelle et perspectiviste du discours.

5 Résultats

L'évaluation finale de notre approche a été réalisée sur l'ensemble de test officiel de la compétition EXIST 2025. Bien que les étiquettes réelles (*gold labels*) demeurent confidentielles, les retours de la compétition nous permettent de situer notre contribution par rapport à l'état de l'art. Pour garantir la fiabilité de nos mesures et évaluer la stabilité de l'architecture, nous présentons les résultats de trois soumissions (*runs*) distinctes, chacune entraînée avec une graine différente (0, 1 et 42). La Table 5 présente nos performances ainsi que notre rang final pour chaque niveau de classification.

Nos résultats sur la classification *soft* sont particulièrement probants : notre modèle s'est classé à la **première place** de la compétition internationale pour cette catégorie, toute tâche confondue. Ce succès valide empiriquement notre architecture hiérarchique `2+1` couplée à l'apprentissage par étiquettes souples. Cette performance démontre aussi que l'intégration du contexte par agents IA

TABLE 5 – Performances globales sur le jeu de test EXIST 2025. Le F1-Score correspond au F1 de la classe OUI pour la Tâche 1, et au Macro-F1 pour les Tâches 2 et 3.

Tâche	Run	Évaluation Soft		Évaluation Hard		
		Rang global	ICM-Soft Norm	Rang global	ICM-Hard Norm	F1-Score
Tâche 1 (Détection)	1	**1** / 71	**0.6700**	16 / 164	0.7878	**0.7802**
	2	2 / 71	0.6690	**15** / 164	**0.7881**	0.7773
	3	3 / 71	0.6662	20 / 164	0.7795	0.7763
Tâche 2 (Intention)	1	**1** / 60	**0.4647**	13 / 144	0.5981	0.5325
	2	2 / 60	0.4592	10 / 144	0.6062	**0.5421**
	3	3 / 60	0.4544	**9** / 144	**0.6117**	0.5384
Tâche 3 (Catégorisation)	1	**1** / 57	**0.4417**	11 / 136	0.5841	**0.6175**
	2	3 / 57	0.4336	16 / 136	0.5657	0.5986
	3	2 / 57	0.4393	**9** / 136	**0.5884**	0.6171

et la modélisation des profils d'annotateurs permettent de capturer avec une précision inédite la distribution de l'incertitude et la diversité des perspectives humaines, métriques clés de l'ICM-Norm. Les performances sur la classification stricte (*hard labels*) reflètent une prudence intrinsèque de l'architecture. En privilégiant la modélisation de l'ambiguïté (probabilités proches de 0,5), le modèle subit une pénalité mathématique lors de la conversion binaire forcée, comparativement aux systèmes optimisés pour la décision unilatérale. Cette dichotomie illustre la tension entre fidélité à la nuance sociologique (*Soft*) et impératif de décision catégorique (*Hard*).

6 Travaux Futurs

Cette étude a démontré l'efficacité d'une approche multitâhce pour la détection du sexisme, articulant l'enrichissement contextuel par agents autonomes, l'intégration des métadonnées sociodémographiques et l'apprentissage par étiquettes souples. Nos résultats confirment que cette synergie est particulièrement performante pour capturer la subjectivité et l'ambiguïté inhérentes au discours haineux, nous permettant d'atteindre la première place dans les catégories de classification souple lors de la campagne EXIST 2025.

Néanmoins, le passage de la probabilité continue à une décision catégorique (*hard labels*) demeure un défi, c'est précisément pour cette raison que notre système se veut un outil d'aide à la décision, sans prétendre à l'autonomie. Nos observations suggèrent que la richesse sémantique capturée par les *soft labels* est partiellement occultée lors du processus de discrétisation nécessaire à l'attribution d'une étiquette unique. Cela souligne la nécessité de développer des stratégies de décision plus sophistiquées, capables de préserver la confiance du modèle dans sa prédiction finale.

Pour l'avenir, afin de réduire la latence induite par la recherche de contexte en temps réel, nous envisageons le développement d'une architecture de mémoire persistante. Une base de connaissances dynamique permettrait de mutualiser les contextes culturels et les tendances émergentes déjà identifiés, ne déclenchant l'agent de recherche web que pour des requêtes inédites.

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.

AMNESTY INTERNATIONAL (2025). Les nouvelles politiques de Meta en matière de contenus risquent d'alimenter davantage de violences de masse et de génocides.

BELBACHIR F., ROUSTAN T. & SOUKANE A. (2024). Detecting online sexism : Integrating sentiment analysis with contextual language models. *AI*, **5**(4), 2852–2863.

CHHABRA A. & VISHWAKARMA D. K. (2023). A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, **29**(3), 1203–1230.

DE VRIES S. & THIERENS D. (2024). Learning with confidence : Training better classifiers from soft labels. *arXiv preprint arXiv :2409.16071*.

DEBNATH A., SUMUKH S., BHAKT N. & GARG K. (2020). Sexist Stereotype Classification on Instagram Data. *URL : https ://github. com/djinn-anthrope/Sexist_Stereotype_Classification*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *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)*, p. 4171–4186.

DONG Z., NI J., BIKEL D., ALFONSECA E., WANG Y., QU C. & ZITOUNI I. (2022). Exploring dual encoder architectures for question answering. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 9414–9419.

FANG Y.-Z., LEE L.-H. & HUANG J.-D. (2024). Nycu-nlp at exist 2024–leveraging transformers with diverse annotations for sexism identification in social networks. *Working Notes of CLEF*.

GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., VAUGHAN A. *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv :2407.21783*.

HAUT CONSEIL À L'EGALITÉ ENTRE LES FEMMES ET LES HOMMES (2026). *Baromètre Sexisme*. Etude 4, Haut Conseil à l'Egalité entre les Femmes et les Hommes.

HE P., GAO J. & CHEN W. (2021). Debertav3 : Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv :2111.09543*.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L., CHEN W. *et al.* (2022). Lora : Low-rank adaptation of large language models. *ICLR*, **1**(2), 3.

INSTITUT DU GENRE EN GÉOPOLITIQUE AND EQUIPOP (2023). *Contrer les discours masculinistes en ligne*. Rapport interne, Institut du Genre en Géopolitique ; Equipop, Paris.

JEONG J. (2023). Hijacking context in large multi-modal models. *arXiv preprint arXiv :2312.07553*.

KIRK H. R., YIN W., VIDGEN B. & RÖTTGER P. (2023). Semeval-2023 task 10 : Explainable detection of online sexism. *arXiv preprint arXiv :2303.04222*.

LEONARDELLI E., UMA A., ABERCROMBIE G., ALMANEA D., BASILE V., FORNACIARI T., PLANK B., RIESER V. & POESIO M. (2023). Semeval-2023 task 11 : Learning with disagreements (lewidi).

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

META (2025). More speech and fewer mistakes.

PENG Y., CHEN Q. & LU Z. (2020). An empirical study of multi-task learning on bert for biomedical text mining. *arXiv preprint arXiv :2005.02799*.

PLAZA L., CARRILLO-DE ALBORNOZ J., AMIGÓ E., GONZALO J., MORANTE R., ROSSO P., SPINA D., CHULVI B., MAESO A. & RUIZ V. (2024). Exist 2024 : sexism identification in social networks and memes. In *Advances in Information Retrieval : 46th European Conference on Information Retrieval*, p. 498–504 : Springer-Verlag.

PLAZA L., CARRILLO-DE ALBORNOZ J., MORANTE R., AMIGÓ E., GONZALO J., SPINA D. & ROSSO P. (2023). Overview of exist 2023–learning with disagreement for sexism identification and characterization. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, p. 316–342.

PLAZA L., DE ALBORNOZ J. C., ARCOS I., ROSSO P., SPINA D., AMIGÓ E., GONZALO J. & MORANTE R. (2025a). Overview of exist 2025 : Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos. In J. C. DE ALBORNOZ, J. GONZALO, L. PLAZA, A. G. S. DE HERRERA, J. MOTHE, F. PIROI, P. ROSSO, D. SPINA, G. FAGGIOLI & N. FERRO, Éds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*.

PLAZA L., DE ALBORNOZ J. C., ARCOS I., ROSSO P., SPINA D., AMIGÓ E., GONZALO J. & MORANTE R. (2025b). Overview of exist 2025 : Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos (extended overview). In G. FAGGIOLI, N. FERRO, P. ROSSO & D. SPINA, Éds., *CLEF 2025 Working Notes*.

QUAN L. M. & THIN D. V. (2024). Sexism identification in social networks with generation-based language models. In *Conference and Labs of the Evaluation Forum*.

RODRÍGUEZ-SÁNCHEZ F., CARRILLO-DE ALBORNOZ J., PLAZA L., GONZALO J., ROSSO P., COMET M. & DONOSO T. (2021). Overview of exist 2021 : sexism identification in social networks. *Procesamiento del Lenguaje Natural*, **67**, 195–207.

RODRÍGUEZ-SÁNCHEZ F., CARRILLO-DE ALBORNOZ J., PLAZA L., MENDIETA-ARAGÓN A., MARCO-REMÓN G., MAKEIENKO M., PLAZA M., GONZALO J., SPINA D. & ROSSO P. (2022). Overview of exist 2022 : sexism identification in social networks. *Procesamiento del Lenguaje Natural*, **69**, 229–240.

SANCHEZ C. & ZHANG Z. (2022). The effects of in-domain corpus size on pre-training bert. *arXiv preprint arXiv :2212.07914*.

STICKLAND A. C. & MURRAY I. (2019). BERT and PALs : Projected attention layers for efficient adaptation in multi-task learning. In K. CHAUDHURI & R. SALAKHUTDINOV, Éds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 de *Proceedings of Machine Learning Research*, p. 5986–5995 : PMLR.

VALAVI E., HESTNESS J., ARDALANI N. & IANSITI M. (2022). Time and the value of data. *arXiv preprint arXiv :2203.09118*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

VETAGIRI A., PAKRAY P. & DAS A. (2025). A deep dive into automated sexism detection using fine-tuned deep learning and large language models. *Engineering Applications of Artificial Intelligence*, **145**, 110167.

WANG L., MA C., FENG X., ZHANG Z., YANG H., ZHANG J., CHEN Z., TANG J., CHEN X., LIN Y. *et al.* (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, **18**(6), 186345.

Annexes

A Prompt Utilisateur pour Recherche de Contexte via Agent IA

Task : Retrieve concise external context to clarify ambiguous tweets or cultural references for sexism classification. Do NOT classify the tweet—only provide context that would help a downstream model to decide.

When to retrieve context :
— The tweet references events, lyrics, memes, or cultural artefacts unfamiliar to a general audience.
— The language is ambiguous (e.g., sarcasm, coded terms, or terms with dual meanings).
— The tweet hints at a broader societal debate or news story.
Guidelines :

1. No classification : Never output YES/NO. Your role is purely contextual.
2. Conciseness : Summarise external context in ≤ 100 tokens.
3. Relevance : Only include context directly tied to potential sexism (e.g., explain a referenced event's sexist controversy, not general info).
4. No context ? Output "No external context needed."

Output Format : [Summary of context, or "No external context needed."]

Examples :

1. **Tweet** : "Ugh, not another 'Boss Babe' anthem. . . "

   Output : "The term 'Boss Babe' is associated with MLM schemes targeting women, often criticised for exploiting feminist rhetoric. Some view it as empowering, others as patronising."

2. **Tweet** : "This is why we need more #NotAllMen energy."

   Output : "#NotAllMen is a hashtag used to critique men who derail conversations about sexism by insisting 'not all men' are problematic. Often cited in debates about systemic misogyny."

3. **Tweet** : "Finally got tickets to the concert !"

   Output : "No external context needed.

B Prompt Utilisateur pour la Tâche 1 - Classification Directs par Agent IA avec accès à Internet

Task : Determine whether a tweet is sexist. Categories : **YES** : The tweet is inherently sexist, describes a sexist situation, or criticises sexist behaviour. Examples :
— “Women are too emotional to hold leadership positions.”
— “At the meeting, all my ideas were ignored until a male colleague repeated them.”
— “Catcalling is not a compliment ; it’s harassment.”
NO : The tweet does not contain sexist content, nor does it describe or criticise sexist situations or behaviours. Examples :
— “Looking forward to the weekend !”
— “Really looking forward to today’s ‘women in web3’ lunch meetup ! If you’re in the la area and want to join, send me a dm ! ! See you ladies soon.”
— “Wow ! Trouble making witches unite !”
Additional Guidelines :
— Ambiguous Language : If the tweet’s sexism is implied rather than explicit, classify it as ’YES.’ If context is insufficient, classify it as ’NO.’
— Strong or Vulgar Language : Classify based on content relevance to sexism, not on the presence of strong language alone.
— Contextual Understanding : Consider societal norms and the broader conversation when evaluating the tweet.
Your final answer will be YES or NO.

Tweet : {tweet}

C Prompt Utilisateur pour la Tâche 1 - Classification LLM avec Contexte

Task : Classify tweets as YES (sexist) or NO (not sexist).

YES : Explicit sexism, descriptions of sexist situations, or criticism of sexism (even implied).

NO : Neutral content. Ignore non-sexist vulgarity. Use societal context.

Answer : (Only YES or NO)

Tweet : {tweet}

Context : {context}

D Configuration du fine-tuning : Hyperparamètres et Prompt (BERT et LLM)

BERT a été fine-tuné avec les hyperparamètres suivants :

— Taux d'apprentissage : 1e-5
— Taille du lot : 64
— Dépréciation du poids : 0,05
— Nombre d'époques : 5

Le LLM a été fine-tuné à l'aide de LoRA, nous avons utilisé la configuration suivante :

— Taille du lot (entraînement et évaluation) : 32
— Étapes d'accumulation du gradient : 4
— Optimiseur : `PagedAdamW_8bit`
— Taux d'apprentissage : 5e-5
— Format de précision : bf16
— Ratio de préchauffage : 0,1
— Rang de décomposition matricielle de LoRA (r) : 4
— Alpha LoRA : 16
— Modules ciblés : `self_attn.q_proj`, `self_attn.k_proj`, `self_attn.v_proj`, `self_attn.o_proj`, `mlp.gate_proj`, `mlp.up_proj` et `mlp.down_proj`

Les prompts zero-shot et du fine-tunage utilisées dans nos expériences sont les suivantes :

> **Task** : Classify tweets as YES (sexist) or NO (not sexist).
>
> **YES** : Explicit sexism, descriptions of sexist situations, or criticism of sexism (even implied).
>
> **NO** : Neutral content. Ignore non-sexist vulgarity. Use societal context.
>
> Answer : (Only YES or NO)
>
> **Tweet :** {tweet}

E Analyse du Contexte Généré

Nous avons procédé à une évaluation préliminaire des contextes générés afin d'examiner leur qualité, leur pertinence et leur exactitude. Notre objectif était d'étudier dans quelle mesure les contextes générés correspondaient aux tweets originaux.

Méthodologie

Nous avons sélectionné au hasard 30 échantillons de contexte dans chaque ensemble de données et les avons évalués selon trois critères :

— **Pertinence** : dans quelle mesure le contexte généré correspondait-il au tweet original ? (Note : 1-5)
— **Exactitude** : le contexte généré fournissait-il des informations ou des aperçus corrects ? (Note : 1-5)
— **Qualité** : le contexte généré était-il cohérent, bien structuré et facile à comprendre ? (Note : 1-5)
— Dans le cas où « aucun contexte externe n'était nécessaire » : était-il approprié de ne pas générer de contexte externe pour le tweet donné ? (Note : 1-5)

Résultats

Cette étude à petite échelle révèle que les contextes générés obtiennent systématiquement des scores parfaits en termes de pertinence (100%) et de qualité. La précision est toutefois satisfaisante, avec

un score moyen de 3,7/5. Il convient de noter que notre modèle démontre une capacité de 100% à identifier les cas où aucun contexte supplémentaire n'est nécessaire. Nous n'avons également observé aucune hallucination dans les textes générés.

Pour approfondir la précision du contexte, nous avons stratifié les résultats en fonction du taux d'accord des six annotateurs sur la classification binaire sexiste du tweet (applicable uniquement aux ensembles de données d'entraînement et de développement, car les résultats des ensembles de données de test ne sont pas disponibles).

TABLE 6 – Analyse de la précision du contexte en fonction du taux d'accord des annotateurs

Taux d'accord des annotateurs	Précision contextuelle
100%	3
83%	4,4
66%	4,3
50%	4,5

Comme le montre la Table 6, nous observons que la précision est moins satisfaisante lorsque le taux d'accord des annotateurs est élevé pour la Tâche 1. Cependant, lorsque les taux d'accord sont plus faibles, la précision tend à s'améliorer. Bien que cette analyse limitée fournisse un premier aperçu encourageant des contextes générés, nous reconnaissons que davantage d'échantillons et d'évaluateurs sont nécessaires pour tirer des conclusions plus solides.