DETECTION DE DISCOURS DE HAINE

Par:
KHENNAOUI Mohamed Seif
SAAD AZZEM Maher Nedjm Eddine
KHERRIB Amina
LEKIKOT Naoufel
NOUICER Sami

Mini-projet WANLP

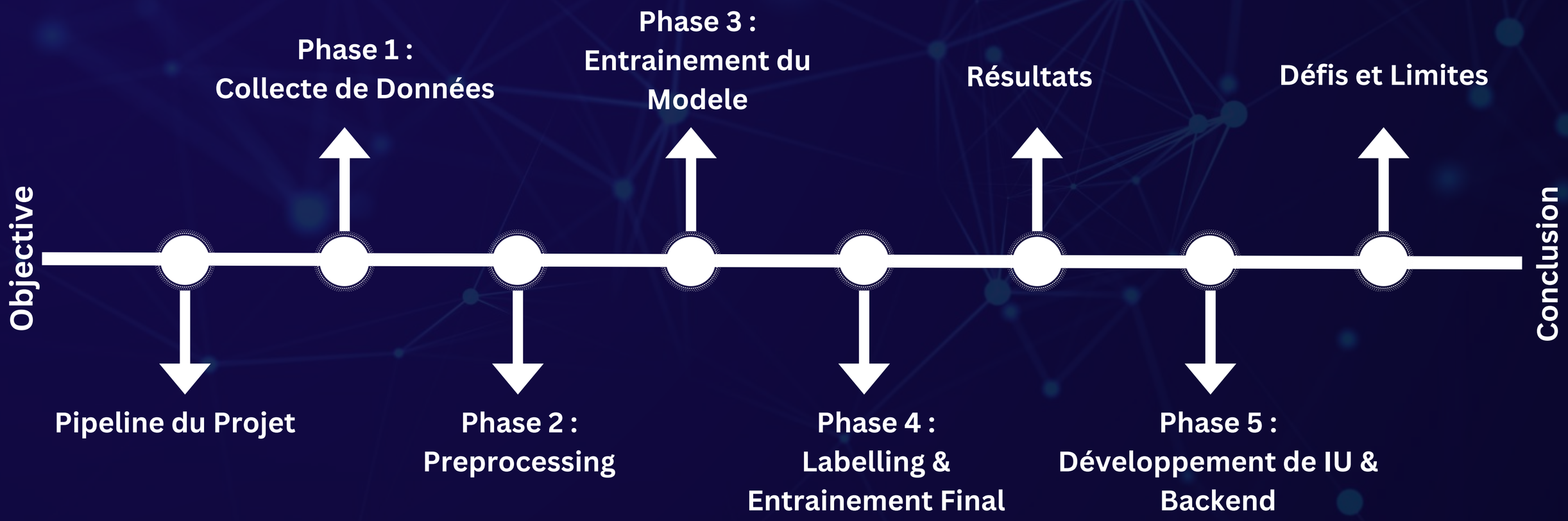
INTRODUCTION

Selon **l'ONU**, un discours de haine est toute communication ou comportement dégradant une personne ou un groupe en raison de son identité (origine, religion, race, genre...).

Selon l'**UNESCO**, les discours de haine **progressent rapidement**, amplifiés par **les réseaux sociaux** !

Face à l'ampleur du phénomène, **la lutte contre les discours haineux** est devenue **une priorité** pour de nombreuses institutions internationales.

Table de Contenu



Objective

La détection de **discours de haine en arabe** est limitée par :

- la complexité de la langue
- la diversité des dialectes.

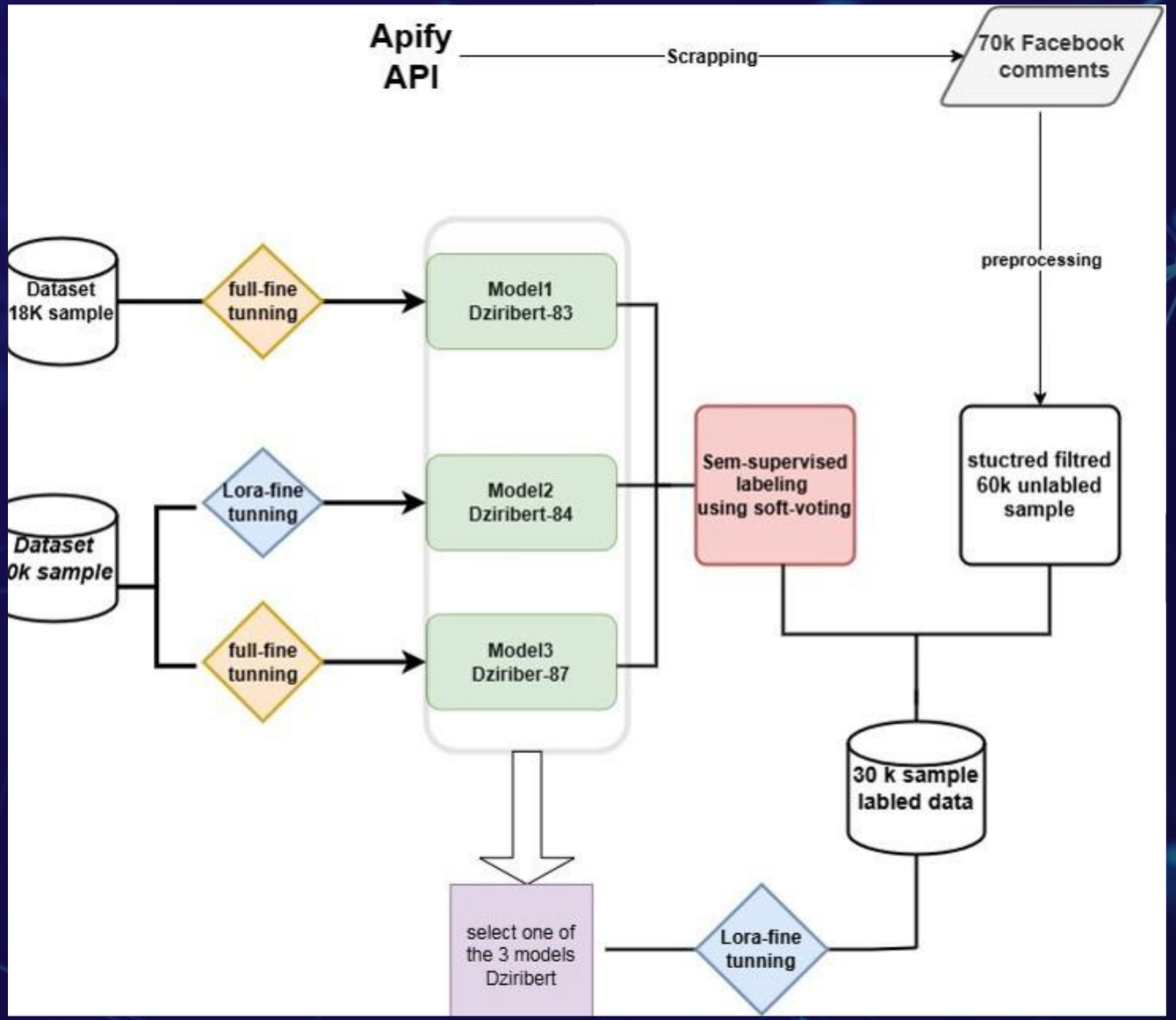
Notre travail vise à développer un modèle capable de **détecter le discours de haine** en **arabe standard** et en **dialecte algérien**, écrit en lettres arabes ou en **Arabizi** (arabe transcrit en alphabet latin).

ÉTAT DE L'ART

Article Consultés

1. Hate speech detection in Algerian dialect using deep learning
2. Deep Learning-based Analysis of Algerian Dialect Dataset Targeted Hate Speech, Offensive Language and Cyberbullying
3. A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere
4. HATE SPEECH DETECTION OF ARABIC SHORTTEXT

Pipeline du Projet



Phase 1 :
Collecte de Données

SOURCES DE DONNEES

1.AlgD_Toxicity_Speech_Dataset
2.Arabizi-Off_Lang_Dataset
3.T-HSAB
4.NArabizi SET
5.Arabic hate speech dataset
6.AJCommentsClassification-
7.GPT-generated
8.Web Scraping (Facebook, YouTube comments)

Phase 1 :
Collecte de Données

DATASET #1 (18K commentaires)

- nous avons fusionné plusieurs jeux de données de discours de haine annotees (hate, non-hate , offensive)

- Ce dataset contient des discours en dialecte algérien + quelques échantillons en dialecte tunisien + marocain

cependant, il convient de souligner que ce dataset présente une **pauvreté notable** en termes de **discours en arabe standard, en Arabizi**

Phase 1 :
Collecte de Données

DATASET #2 (40K commentaires)

Nous devions enrichir notre jeu de données avec plus de texte en Arabe standard et en Arabizi. Finalement, nous avons obtenu un dataset de 40K+ commentaires.

DATASET #1

18K commentaire

extraits de 5 différents

datasets annotés :

haine/non-haine

DATASET #2

Dataset#1

+

records from NArabizi dataset

+

GPT-generated data

=

40K+

Phase 1 :
Collecte de Données

DATASET #3 (60K + commentaires)

➤ **WEB SCRAPING using Apify** (Facebook et YouTube)

un dataset plus riche en termes de:

- **diversité des catégories** de discours de haine
- **thématiques** abordées

Phase 1 :
Collecte de Données

DATASET #3 (60K + commentaires)

➤ **WEB SCRAPING using Apify** (Facebook et YouTube)

un dataset plus riche en termes de:

- **diversité des catégories** de discours de haine
- **thématiques** abordées

facebook comments
youtube comments
commentaire sur des
article d'aljazeera

Comment annoter plus de 60 K commentaires

Phase 1 :
Collecte de Données

DATASET #3 (60K + commentaires)

WEB SCRAPING using Apify (Facebook et YouTube)

un dataset plus riche en termes de:

- **diversité des catégories** de discours de haine
- **thématiques** abordées

facebook comments
youtube comments
commentaire sur des
article d'aljazeera

Phase 2 : Preprocessing

Supprimer les noms d'utilisateur et les URL

Normaliser la casse et les caractères arabes

Normaliser les espaces

Phase 2 : Preprocessing

- Éliminer les @mentions
- Retirer les liens http://exemple.com
- Supprimer les espaces inutiles en début et fin de texte.
- Convertir le texte en minuscules et standardiser les variantes des lettres arabes (ex. إ, أ, آ, ا en ا).

Phase 3 : Entrainement du Modele

Model #1
DziriBERT-83

- Données : 18 000 échantillons validés (Arabizi + dialecte algérien, et arabe standard)
- Méthode : Fine-tuning complet

Phase 3 : Entrainement du Modele

Model #2
DziriBERT-84

- Données : 40 000 échantillons validés (Arabizi, dialecte algérien et arabe standard), augmentés avec GPT-texte pour équilibrer les classes.
- Méthode : Fine-tuning avec lora

Phase 3 : Entrainement du Modele

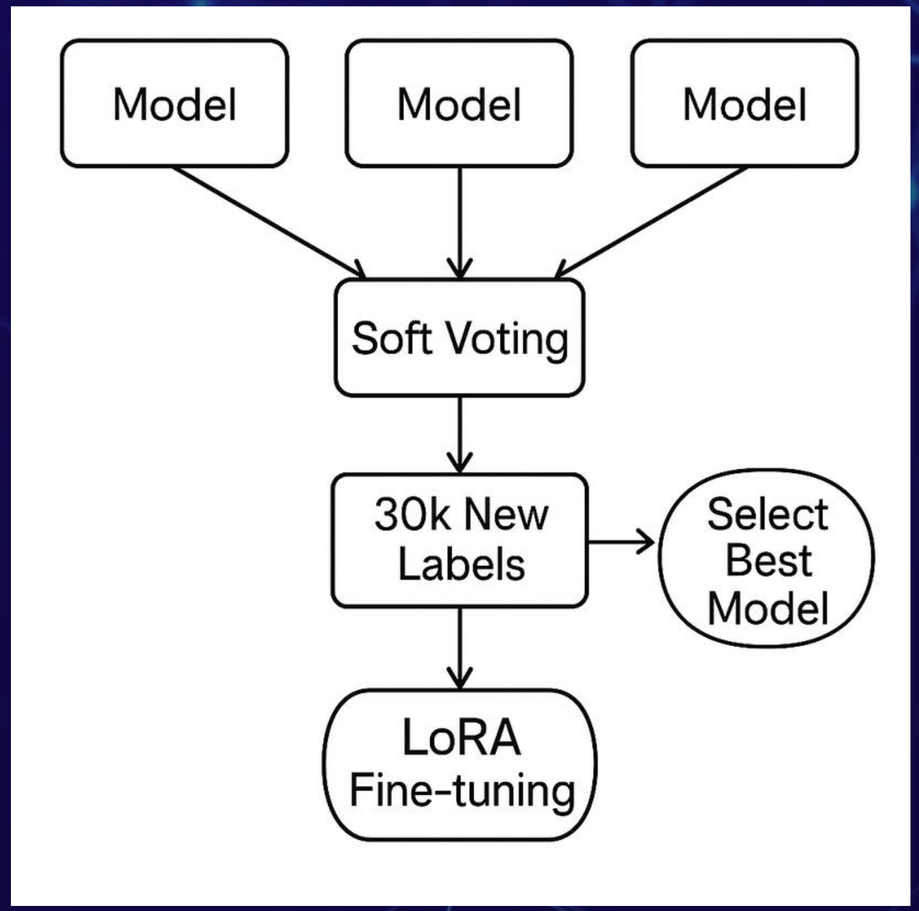
Model #3
DziriBERT-87

- Données : 40 000 échantillons validés (Arabizi, dialecte algérien et arabe standard), augmentés avec GPT-texte pour équilibrer les classes.
- Méthode : Fine-tuning complet

Phase 3 : Entrainement du Modele

METRIQUE	MODEL1 (dziribert-83)	MODEL2 (dziribert-84)	MODEL3 (Dziribert-87)
ACC	0.83	0.8425	0.8744
F1	0.8331	0.8444	0.8743
PRECISION	0.833	0.8341	0.8750
RECALL	0.833	0.8550	0.876

Phase 4 :
Labelling &
Entrainement Final

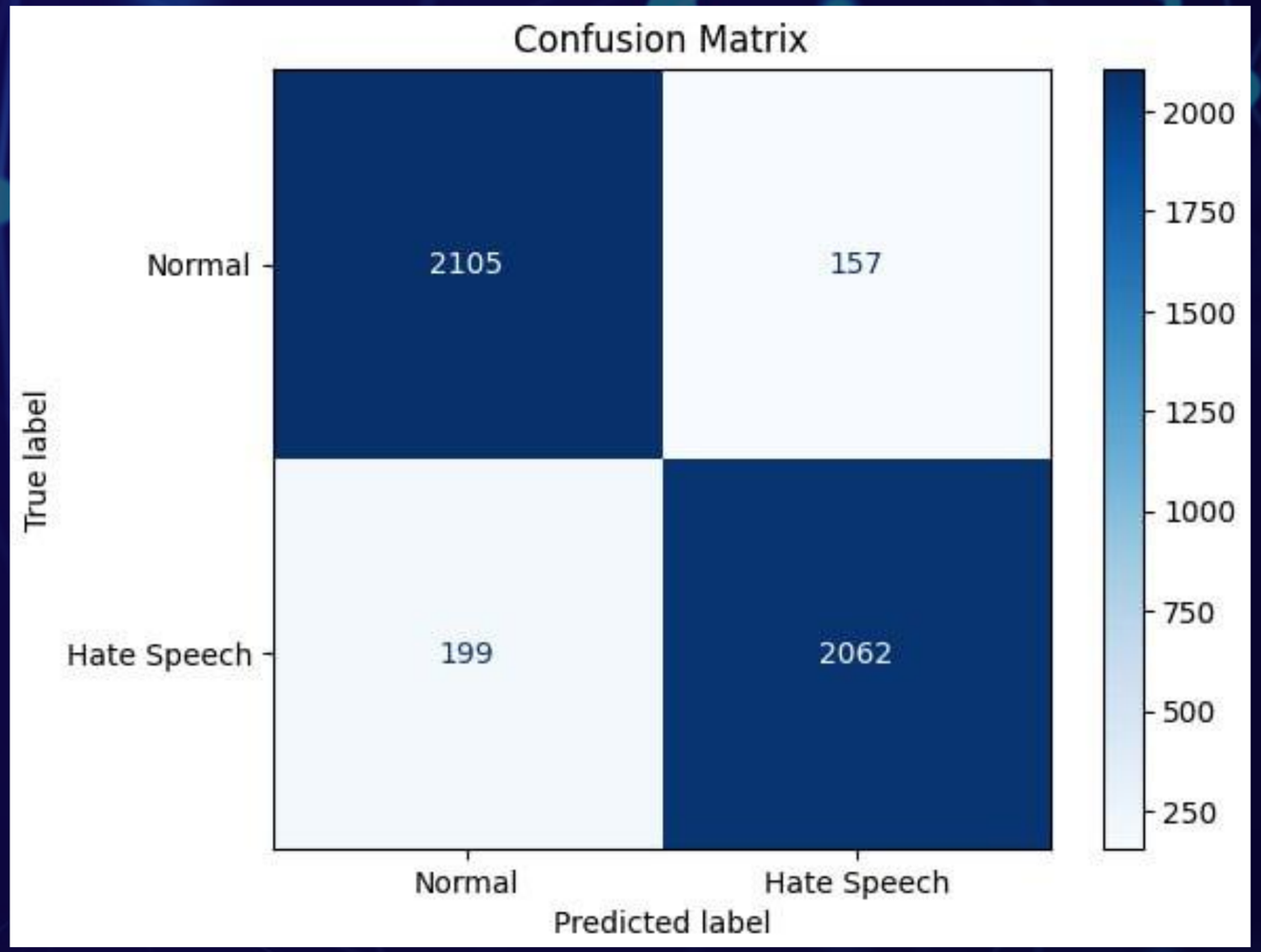


- Utilisation de l’apprentissage Semi Supervisé pour l’annotation
- Annotation de nouvelles données par soft voting (moyenne pondérée des prédictions).
- 30 000 nouvelles exemples fiables générés pour enrichir notre dataset.
- Sélection du meilleur modèle puis affinage final avec LoRA fine-tuning.

Résultats

👉 Confusion Matrix

- Précision et Recall équilbrés (92%)
- Accuracy plus elevé (93%)

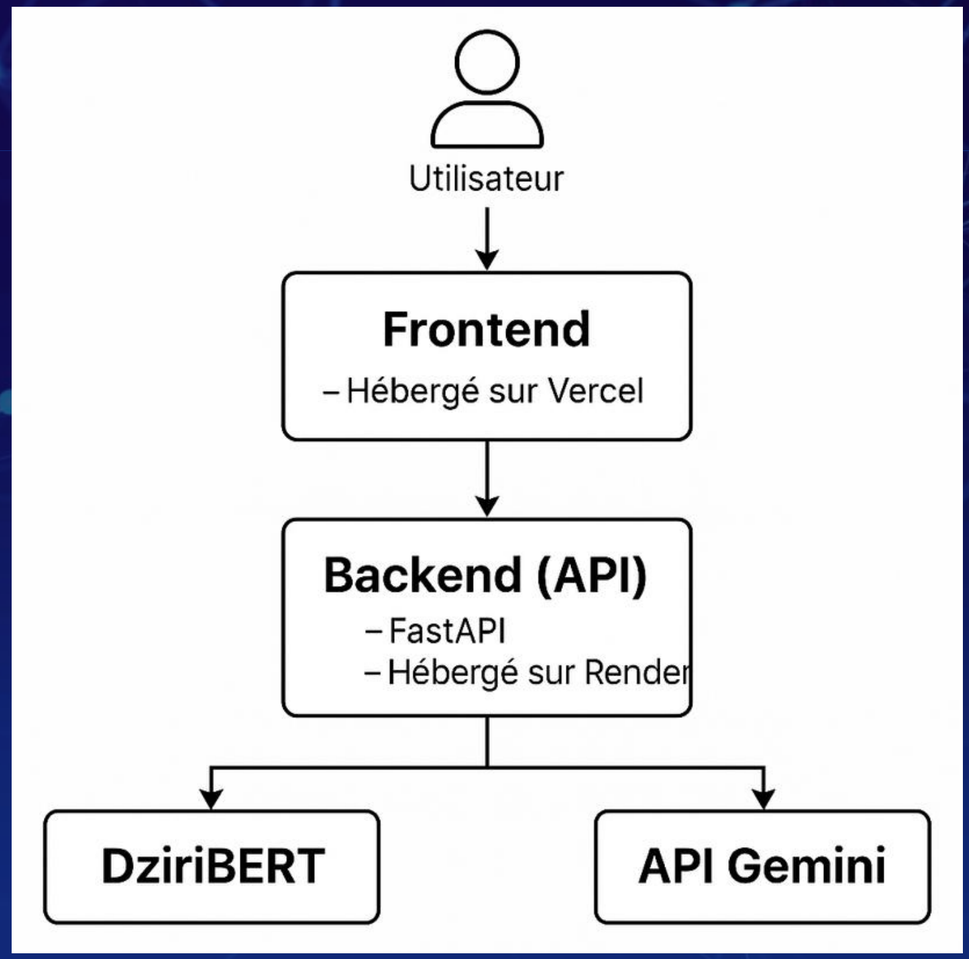


Résultats

🔵 OLD MODEL PREDICTIONS:
Text: الرجل أعلى شأنا من المرأة ---→ Prediction: ✅ Normal
Text: la souffrace m3a l3rab mou7al tokhlos ---→ Prediction: ✅ Normal
Text: لا يرتقون للبويض lk7alach ---→ Prediction: ✅ Normal
Text: الآسيويون هم اسباب كل امراض العالم ---→ Prediction: 🚨 HATE SPEECH
Text: l9bayel bara men dz ---→ Prediction: ✅ Normal
Text: احب العيش في الذل ---→ Prediction: ✅ Normal

🔴 LoRA MODEL PREDICTIONS:
Text: الرجل أعلى شأنا من المرأة ---→ Prediction: 🚨 HATE SPEECH
Text: la souffrace m3a l3rab mou7al tokhlos ---→ Prediction: 🚨 HATE SPEECH
Text: لا يرتقون للبويض lk7alach ---→ Prediction: 🚨 HATE SPEECH
Text: الآسيويون هم اسباب كل امراض العالم ---→ Prediction: 🚨 HATE SPEECH
Text: l9bayel bara men dz ---→ Prediction: 🚨 HATE SPEECH
Text: احب العيش في الذل ---→ Prediction: 🚨 HATE SPEECH

Phase 5 :
Développement de IU & Backend



Phase 5 : Développement de IU & Backend

Interface Utilisateur Frontend (React.js + MUI)

- Interface utilisateur légère et simple
- Conçue pour permettre aux utilisateurs de :
  - Saisir un texte
  - Voir le résultat : Non-haineux ou Haineux
- Si haineux, afficher des informations supplémentaires : Catégorie et Description
- La communication se fait via des appels API vers le backend.

Phase 5 : Développement de IU & Backend

Modèles d'IA Utilisés

API Gemini :

- Génère la catégorie du discours haineux (ex : racisme, religion)
- Décrit le contexte du discours haineux dans un texte compréhensible par l'humain

DziriBERT :

- Modèle linguistique pré-entraîné pour l'arabe algérien
- Modèle léger, servi localement via le backend

Phase 5 : Développement de IU & Backend

Implémentation du Backend (FastAPI)

- Le framework FastAPI est utilisé pour sa rapidité et son support asynchrone facile.
- FastAPI gère :
  - La classification de texte (haine / non-haine).
  - Si un discours haineux est détecté → Appel à l'API Gemini pour :
    - Obtenir catégorie.
    - Rédiger une courte description.
- Programmation asynchrone pour des réponses plus rapides.

Défis et Limites

Collecte et Préparation des Données

- Trouver ou construire un dataset adapté à l’algérien (dz) et au discours haineux.
- Mélanger deux types de données (discours haineux vs discours normal).
- Nettoyage et pré-traitement des données (retirer le bruit, gérer les textes courts/incomplets).
- Gérer le déséquilibre de classes (non-hate, hate) et biais vers quelques categories .
- Préparation d'un jeu de données de qualité,

Défis et Limites

Optimisation de la Rapidité et de la Stabilité:

- Assurer un temps de réponse rapide malgré l’appel à deux modèles différents (DziriBERT + Gemini).
- Héberger le Frontend sur Vercel et le Backend sur Render demande de gérer des contraintes de compatibilité, d'environnement et de performance.

CONCLUSION

- Un système de détection automatique des discours de haine en arabe standard, dialecte algérien et Arabizi a été développé.
- L'approche adoptée traite la complexité linguistique et la diversité des dialectes à travers un pipeline complet basé sur des techniques de deep learning.
- Ce travail constitue une base solide pour des recherches futures visant à améliorer la qualité des données et l'efficacité des modèles.

MERCI POUR VOTRE ATTENTION !