Projet “Hate Speech”
Sur la base de l’article “Détection multilabel de discours haineux et de langage abusif dans le Twitter indonésien”

Muhammad Okky Ibrohim
Indra Budi
Faculté d'informatique Universitas Indonesia

Mikhail BIRIUCHINSKII
Alina MIASNIKOVA

2023-2024

Université Paris Nanterre

Table de matières

Dépôt Github :

https://github.com/MichaBiriuchinskii/Hate_Speech_Project

1) Introduction de projet

L’importance du sujet de l'étude :

- **Adaptation culturelle**
- **Annotation linguistique**
- **Apprentissage automatique**

2) Présentation de l’article

Définitions

Discours de haine - un discours direct ou indirect envers une personne ou un groupe contenant de la haine fondée sur quelque chose d'intrinsèque à cette personne ou ce groupe.

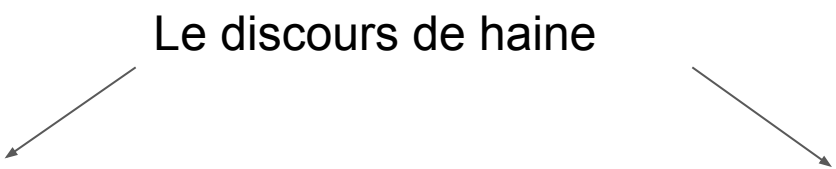
"@detikcom Demandez à Amin Rais... Autrefois, c'est lui qui a semé le trouble et a appelé à la réforme. Les élections locales en sont l'un des résultats."*

Langage abusif - une énonciation qui contient des mots/phrases offensifs qui est communiquée à l'interlocuteur (individus ou groupes), aussi bien verbalement qu'à l'écrit.

"@junichirich @detikcom Voici les effets de l'enseignement des partisans de la secte cingkrang, avec leur front bas et leurs fesses en forme de casserole."

*À partir d'ici, sont fournis des traductions de l'indonésien réalisées avec l'aide de Google Translate et de notre humble intuition de traducteur

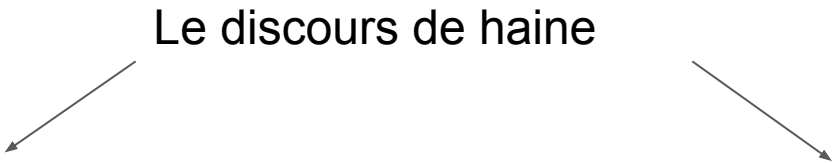
2) Discours de haine : cibles, catégories et niveaux en Indonésie



"Ramos, que je considérais autrefois comme mon idole, je le trouve maintenant répugnant, maudit soit-il."

"@Giring_Ganesha PSI EST UN PARTI ADMIRATEUR DE DICTATEURS.."

2) Discours de haine : cibles, catégories et niveaux en Indonésie



individus

"Ramos, que je considérais autrefois comme mon idole, je le trouve maintenant répugnant, maudit soit-il."

groupes

"@Giring_Ganesha PSI EST UN PARTI ADMIRATEUR DE DICTATEURS.."

2) Selon les résultats du FG

Les catégories de discours de haine

1. **Religion/croyance**
2. **Race/ethnie**
3. **Physique/handicap**
4. **Genre/orientation sexuelle**
5. **Autre calomnie**

2) Selon les résultats du FG

Les niveaux de la haine :

1. **Le discours de haine faible**
   “ @MafiaWasit Vos propos sont déplacés… espèce de [...] ”

2. **Le discours de haine modéré**
   “Les chrétiens sont les mêmes que les colonisateurs, c'est dangereux !!!”

3. **Le discours de haine forte**
   ”Jokowi et ses partisans ne sont que des laquais chinois qui causent des problèmes à la nation ! Boycottez tous les programmes de Jokowi et expulsez tous les agents chinois pour le progrès de la nation !”

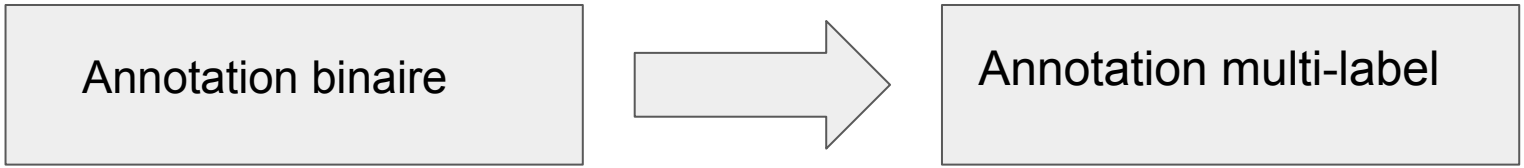
2) Collecte de données et annotation

Dataset

Langue : indonésien

13 169 tweets

2) Collecte de données et annotation



Critères de sélection des annotateurs

- Avoir entre 20 et 30 ans, car la plupart des utilisateurs de Twitter ont cet âge en Indonésie.
- Être locuteur natif de l'indonésien.
- Être un utilisateur expérimenté de Twitter.
- Ne pas être affilié à une organisation ou un parti politique.

2) Résultats d'annotation

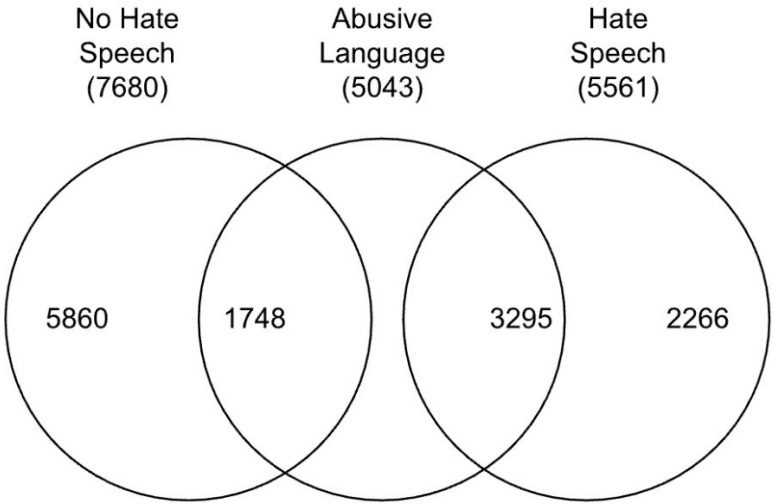
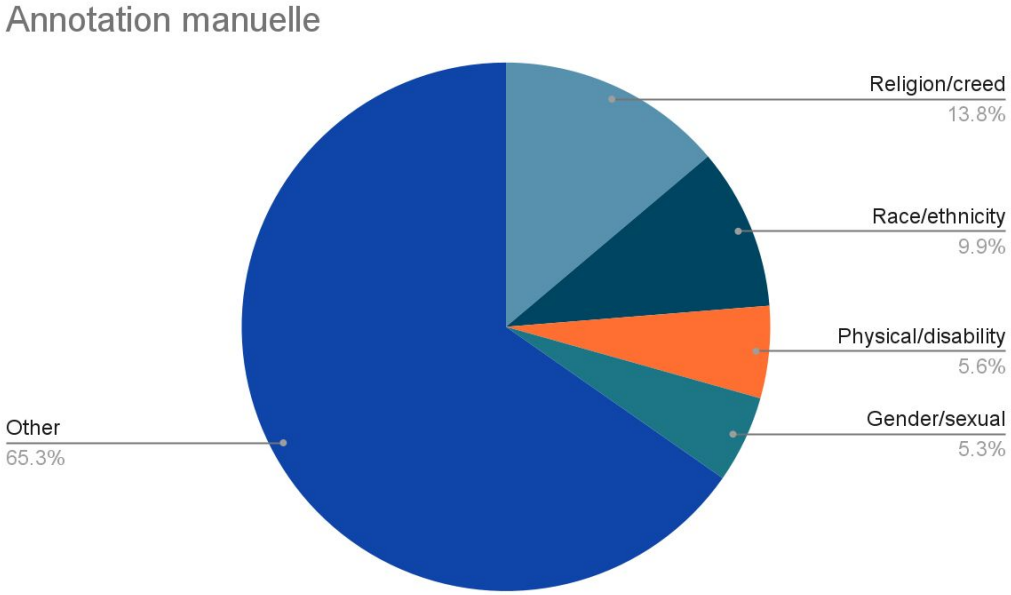


Figure 1 : Répartition du langage abusif entre les tweets ne contenant pas de propos haineux et les tweets contenant des propos haineux

2) Résultats d’annotation



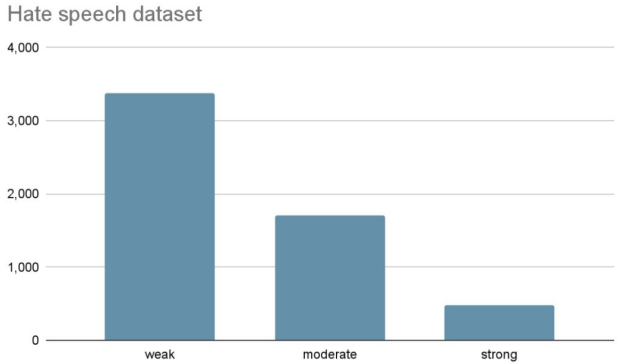


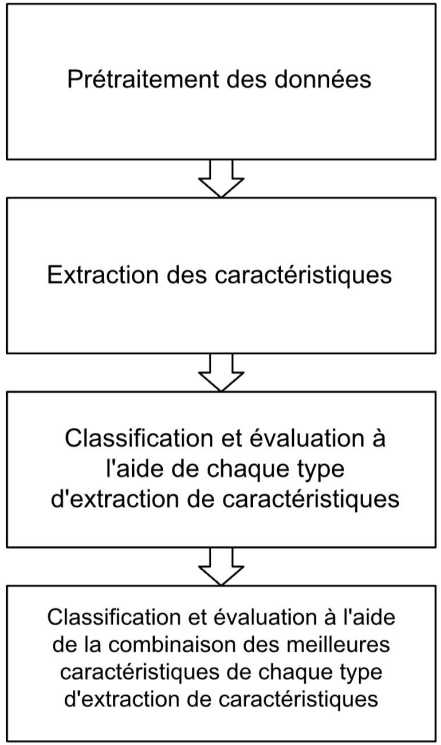
Figure 3 : La répartition des niveaux de la haine

2) Expériences

L'expérience a 2 scénarios :

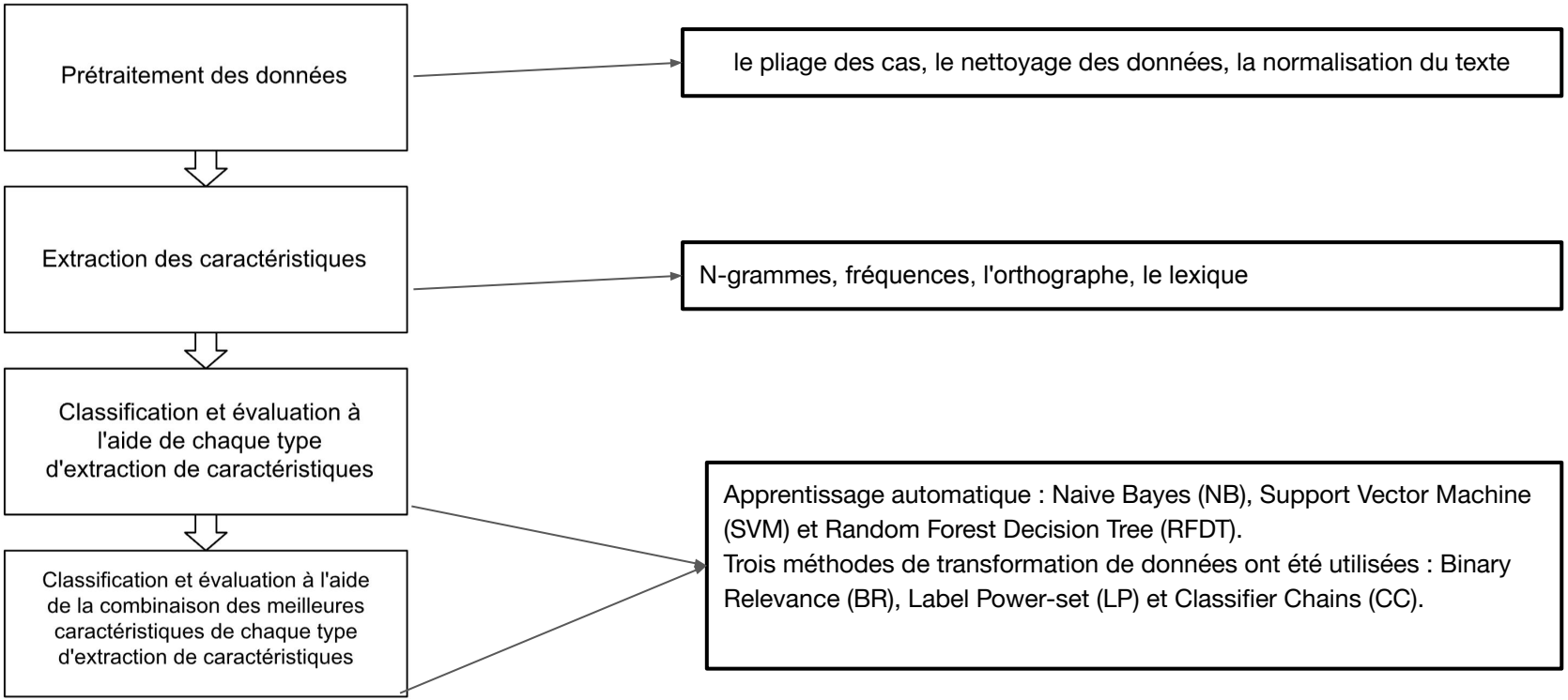
Le premier - d'identifier le langage abusif et les discours de haine, **y compris** la cible, les catégories et le niveau qui sont contenus dans un tweet.

Le deuxième, c'est la classification multi-label pour identifier le langage abusif et les discours de haine contenus dans un tweet **sans** identifier la cible, les catégories et le niveau de discours de haine.



La chaîne de traitement

2) Expériences



La chaîne de traitement

2) Résultats de la première expérience

Type de caractéristique	Meilleure caractéristique sur la base de la précision moyenne	Précision moyenne (%)
n-grammes de mots	mot unigramme + bigramme + trigramme	59.44
n-grammes de caractères	caractère quadgrammes	52.55
ortographie	point d'interrogation	44.44
lexique	negative sentiment	44.45

Meilleurs classifieurs : RFDT, LP

2) Résultats du deuxième expérience

Type de caractéristique	Meilleure caractéristique sur la base de la précision moyenne	Précision moyenne (%)
n-grammes de mots	mot unigramme + bigramme + trigramme	73.53
n-grammes de caractères	caractère quadgrammes	72.44
ortographie	point d'exclamation	45.27
lexique	sentiment positif + lexique abusif	52.10

Tableau 2 : Meilleure performance de chaque type d'extraction de caractéristiques sur la base de la précision moyenne pour le deuxième scénario

Meilleurs classifieurs : Le classificateur RFDT avec la méthode de transformation de données LP

3) Mise en pratique du modèle linguistique élaboré par les auteurs

Hugging Face

193 tweets en russe

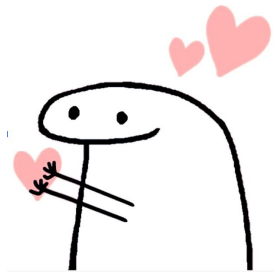
"Il boit chaque semaine et cela ne le rend pas en meilleure santé."

"La crise de la masculinité n'est en réalité pas une crise."

"Je viendrai à votre rassemblement avec des chaînes et vous ferai tous fuir."

*"Мы с братанами на ауешном движении, мы настоящие волки в законе"**

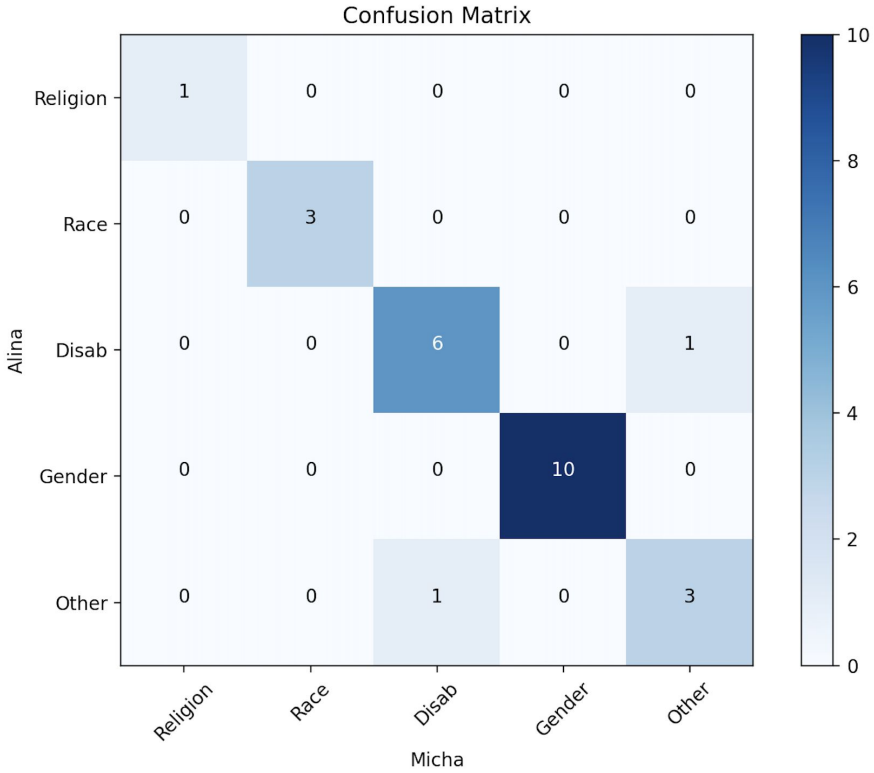
**bonus pour les russophones*



3) Évaluation de l'annotation

Kappa (K) = (Po - Pe) / (1 - Pe)

(0.92 - 0,00017203) / (1 - 0,00017203) = 0,91998624



3) La pertinence des expériences d'apprentissage automatique

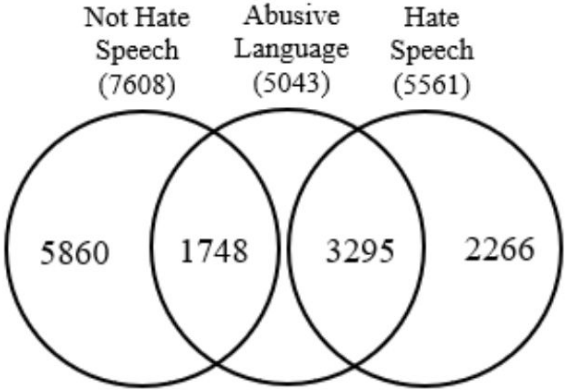
```
Accuracy: 0.717948717948718
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.92      0.81        26
           1       0.67      0.31      0.42        13

    accuracy                           0.72        39
   macro avg       0.70      0.62      0.62        39
weighted avg       0.71      0.72      0.68        39
```

SVM classifier sur la base ‘Hate speech’ ou ‘No hate speech’

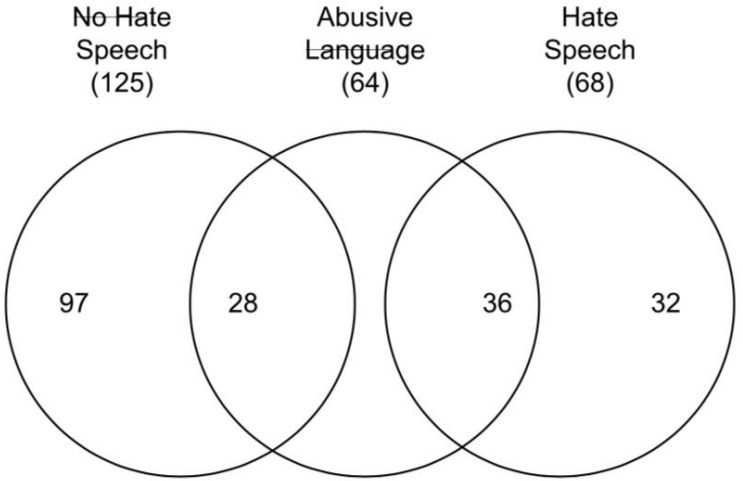
Découpage en unigrammes et bigrammes

Vectorisation avec TfidfVectorizer, entraînement avec SVC(kernel='linear')

3) Mise en pratique du modèle linguistique élaboré par les auteurs

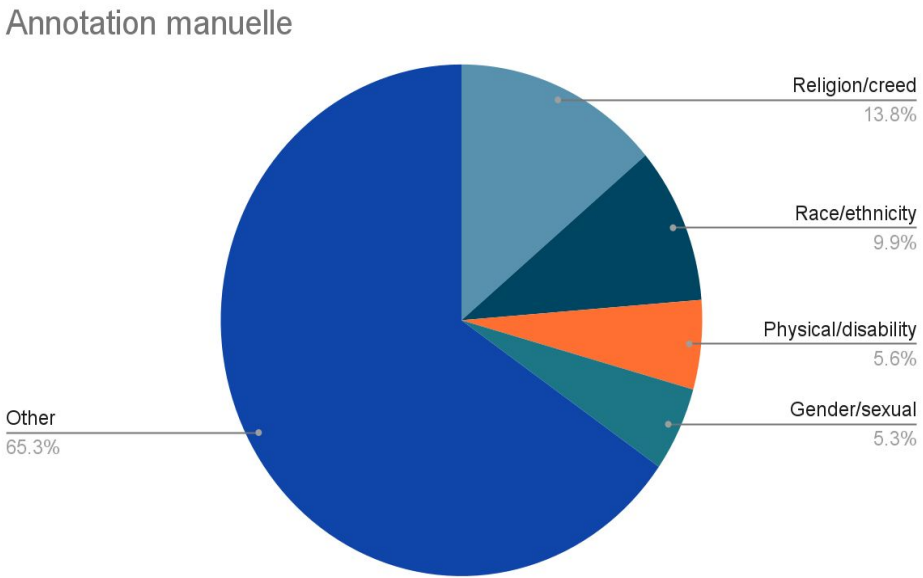


Dataset en indonésien

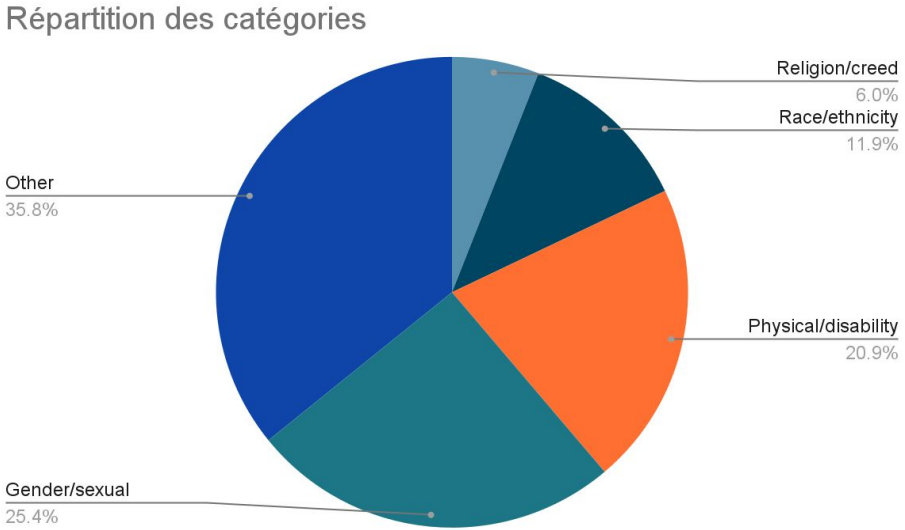


Dataset en russe

3) Mise en pratique du modèle linguistique élaboré par les auteurs

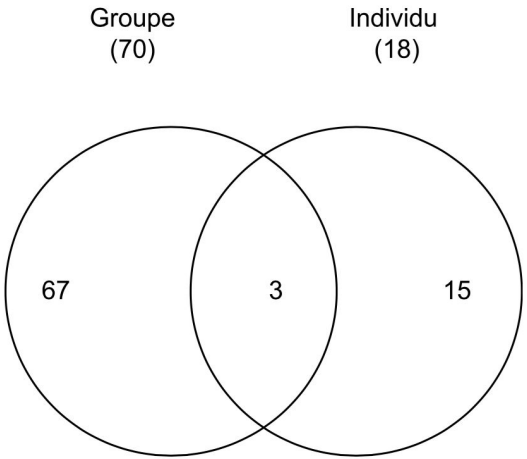


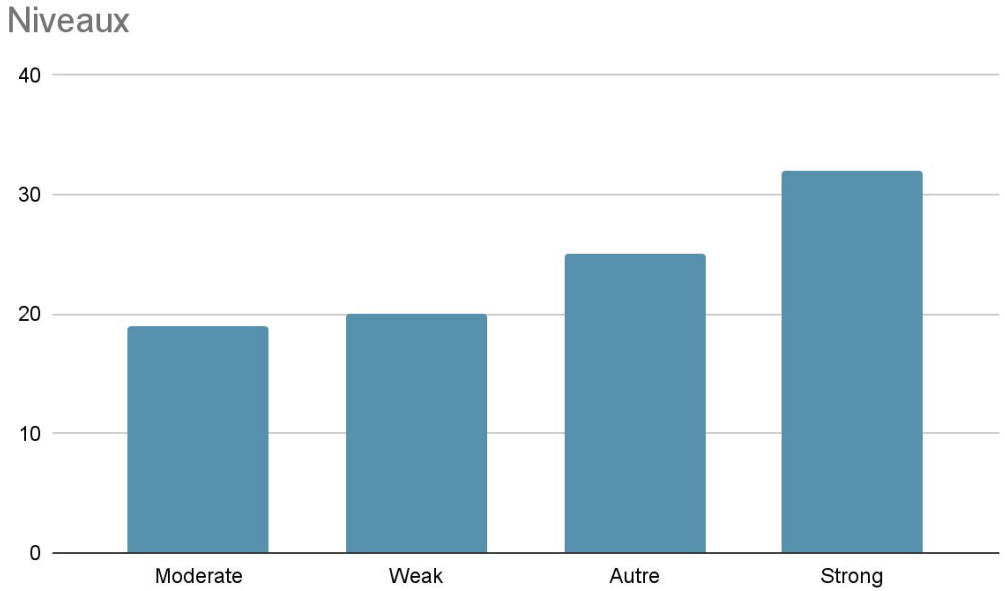
Dataset en indonésien



Dataset en russe

3) Mise en pratique du modèle linguistique élaboré par les auteurs





3) Conclusion

Tendances similaires :

La plupart des messages sont *no-hate* mais abusifs

La prédominance de la catégorie 'autre'

La partie Machine Learning reste difficile mais prometteuse

Tendances différentes :

Le ciblage de groupes plutôt que d'individus

Suggéstions :

Adapter les catégories

Économiser les ressources (les features non-utilisés)

Traitement des langues moins fréquentes

Exploration des modèles de deep learning

Merci

Merci