De la modélisation au traitement automatique des données linguistiques

Projet “Hate speech”

Réalisé à la base de l’article scientifique :
“Détection multilabel de discours haineux et de langage abusif dans le Twitter indonésien”

Mikhail BIRIUCHINSKII
Alina MIASNIKOVA

2023-2024

Université
Paris Nanterre

Alina M. Mikhail B.

Table de matières

Alina M. Mikhail B.

1) Introduction de projet

Le présent rapport de projet vise à présenter notre démarche dans le cadre d'une recherche visant à adapter et étendre le projet linguistique de l'article intitulé "Détection multilabel de discours haineux et de langage abusif dans le Twitter indonésien[1]".

Notre objectif principal est d'appliquer l'annotation linguistique développée dans l'article à un corpus de données en langue russe afin d'évaluer la pertinence du modèle et de développer notre propre compréhension de telles tâches de traitement automatique des langues.

Le discours haineux et le langage abusif en ligne sont des problèmes croissants qui affectent la qualité de la communication numérique et ont des répercussions significatives sur la société. Ces comportements nuisibles sont omniprésents sur les réseaux sociaux, forums en ligne et autres plateformes, et ils peuvent entraîner des conséquences graves, telles que la diffusion de la haine, la discrimination, la polarisation et la violence. Il est impératif de lutter contre ces formes de discours afin de préserver un environnement en ligne plus sain et respectueux.

La recherche sur la détection de discours haineux et de langage abusif est essentielle pour relever ce défi. Elle peut être utilisée pour automatiser le processus de surveillance et de modération des contenus en ligne, identifiant ainsi rapidement et efficacement les discours problématiques. Cette automatisation est particulièrement pertinente compte tenu de la volumétrie considérable des données générées en ligne chaque jour.

Le choix de l'article susmentionné ainsi que l'importance du sujet de l'étude s'explique, de notre point de vue, par les facteurs suivants :

Adaptation culturelle : Alors que l'article initial se concentrait sur le Twitter indonésien, notre projet s'adresse à un contexte linguistique et culturel différent, à savoir le russe. Cette adaptation est essentielle car la perception de ce qui constitue un discours haineux ou abusif peut varier d'une culture à l'autre.

Annotation linguistique : Le russe est l'une des langues les plus parlées au monde, mais ce n'est pas la langue la plus dotée en termes d'outils et de recherche en matière de traitement automatique des langues. Il est donc très important de développer des modèles linguistiques capables de gérer différents défis, tels que la détection de la haine avec le langage souvent familier.

[1] https://aclanthology.org/W19-3506/

Apprentissage automatique : L'article étudié s'appuie sur des techniques d'apprentissage automatique pour entraîner des modèles de détection de discours haineux. Il nous était intéressant de tester cet aspect, pour voir si cela vaut le coup d'essayer les outils de la classification automatiques avec notre jeu de données ainsi qu'avec nos ressources.

Tout au long des pages suivantes, nous allons tout d'abord vous donner un bref résumé du document original sur lequel nous avons basé l'expérience. Ensuite, nous présenterons une partie du travail effectué avec nos données, nous parlerons du bien-fondé de l'utilisation d'outils d'apprentissage automatique dans notre cas, et nous conclurons.

Tous les matériaux tels que les données, les annotations et les références peuvent être trouvés sur le dépôt préparé à cet effet[2].

2) Présentation de l'article scientifique

- **Définitions élaborées par les auteurs**

Discours de haine est 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 (Komnas HAM, 2015 - une institution indépendante chargée de mener des études, des recherches dans le domaine des droits de l'homme en Indonésie).

"@detikcom Tanya amin rais.. Dulu dia yg bikin kacau dan nyuruh reformasi. Pilkada salah satu hasilnya."
Traduction : "@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."

Le **langage abusif** est une énonciation qui contient des mots/phrases offensifs qui est communiquée à l'interlocuteur (individus ou groupes), aussi bien verbalement qu'à l'écrit. Par conséquent, le discours de haine peut contenir un langage abusif.

"@junichirich @detikcom Inilah efek ajaran bani cingkrang jidat item pantat panci."
Traduction : "@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."

Le travail est unique dans le sens où les chercheurs ont remarqué que dans les travaux similaires la classification n'est pas suffisamment affinée. Par exemple, on annote uniquement le degré de haine (fort, moyen, faible), voire note tout simplement si le texte contient du discours de haine ou pas. Les auteurs ont donc décidé de faire, dans un premier temps, la distinction entre le discours de haine et le langage abusif (ne sont pas

[2] https://github.com/MichaBiriuchinskii/Hate_Speech_Project

mutuellement exclusifs). Pour classifier le discours de haine, les chercheurs ont introduit des critères supplémentaires : cible, catégorie et degré du discours de haine.

Le dataset a été recueilli à partir des tweets en indonésien à l'aide de l'API de recherche Twitter, qui est mise en œuvre à l'aide de la bibliothèque Tweepy. Twitter a été choisi parce que Twitter est l'une des plateformes de médias sociaux en Indonésie qui est souvent utilisée pour diffuser des discours haineux et des propos injurieux d'après les auteurs.

Cette recherche traite de la classification de texte multilabel pour la détection de langage abusif et de discours haineux, y compris la détection de la cible, de la catégorie et du niveau de discours haineux dans le Twitter indonésien en utilisant des approches d'apprentissage automatique avec la machine à vecteur de support (SVM), Naive Bayes (NB) et l'arbre de décision Random Forest (RFDT) et la pertinence binaire (BR), l'ensemble d'étiquettes (LP) et les chaînes de classificateurs (CC) comme méthode de transformation des données. Les chercheurs ont utilisé plusieurs types d'extraction de caractéristiques, à savoir la fréquence des termes, l'orthographe et les caractéristiques du lexique.

- **Discours de haine : cibles, catégories et niveaux en Indonésie**

Dans cette étude, un focus groupe (**FG**) a été constitué pour interroger les participants sur les cybercrimes en Indonésie. L'objectif était d'obtenir une définition précise de la haine en ligne, y compris ses caractéristiques. Le focus groupe a révélé que la haine en ligne cible spécifiquement des *individus ou des groupes* qui ont des *catégories* et des *niveaux* particuliers.

Le discours de haine à cible individuelle vise une personne spécifique, tandis que le discours de haine à cible collective vise des groupes, des associations ou des communautés particulières.

Cible individuel : "Ramos yang aku pandang idola dahulu, sekarang dah ku anggap jijik babi kmk"
Traduction : "Ramos, que je considérais autrefois comme mon idole, je le trouve maintenant répugnant, maudit soit-il."

Cible collective : "@Giring_Ganesha PSI ITU PARTAI PENGAGUM DIKTATOR.."
Traduction : "@Giring_Ganesha PSI EST UN PARTI ADMIRATEUR DE DICTATEURS.."

Ces groupes, associations et communautés peuvent prendre la forme de groupes religieux, de races, de tendances politiques, de clubs de fans, de communautés de loisirs, etc. Que ce soit dirigé vers une personne ou un groupe, le discours de haine repose sur une catégorie particulière.

Selon les résultats du FG, en général, les catégories de discours de haine sont les suivantes :

1. **Religion/croyance**, correspond à un discours de haine basé sur une religion (islam, christianisme, catholicisme, etc.) ou une croyance particulière ;

2. **Race/ethnie**, englobe un discours de haine basé sur des caractéristiques physiques telles que la forme du visage, la taille, la couleur de la peau, etc. ou sur l'ethnie ;

3. **Physique/handicap**, est un discours de haine basé sur des déficiences physiques (par exemple, la forme du visage, des yeux et d'autres parties du corps) ou sur des handicaps (par exemple, l'autisme, l'idiotie, la cécité, la surdité, etc.) ;

4. **Genre/orientation sexuelle**, concerne un discours de haine basé sur le genre (homme et femme), qui implique les mots offensifs (par exemple : gigolo, salope, etc.) ou sur une orientation sexuelle non conforme (par exemple : homosexuel, lesbienne, etc.) ;

5. **Autre calomnie**, englobe les propos injurieux qui ne font pas partie des catégories décrites ci-dessus.

Outre les cibles et les catégories, le discours de haine a également un certain niveau. Sur la base des résultats du focus group, le discours de haine peut avoir l'un des trois niveaux : faible, modéré et fort.

1. Le **discours de haine faible** consiste en des commentaires négatifs sur des personnes, mais qui n'incitent pas au conflit.

   " @MafiaWasit Omongan mu rusak.. Dasar KontoL "
   " @MafiaWasit Vos propos sont déplacés… espèce de [...] "

2. Le **discours de haine modéré** consiste en des commentaires négatifs visant un groupe de personnes, et pouvant conduire à un conflit qui se limiterait à des échanges sur les réseaux sociaux.

   "Kristen itu sama dengan penjajah, bahaya itu!!"
   "Les chrétiens sont les mêmes que les colonisateurs, c'est dangereux !!!"

3. Le **discours de haine forte** vise des individus ou des groupes de personnes et doit être considéré comme prioritaire parce qu'il peut conduire à des conflits réels et physiques.

“Jokowi dan para cebongnya hanyalah antek cina yang menyusahkan bangsa! Boikot semua program Jokowi dan usir semua antek Cina demi kemajuan Bangsa!1!1!1”
”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 !”

- **Data Collection and Annotation**

Pour cette recherche les auteurs ont constitué le corpus avec les tweets tirés à l'aide de la bibliothèque Tweepy ainsi que l’ensemble de données sur les discours de haine et les propos injurieux provenant de plusieurs recherches antérieures.

Le processus d'annotation a été effectué par des non-linguistes et a été divisé en deux parties : l'annotation des propos dits "haineux/injurieux" (16,500 tweets) et l'annotation des propos en fonction de leur type de haine et de leur niveau (5,700 tweets). Cependant, pour la phase de préparation, les chercheurs ont eu plusieurs consultations avec un linguiste afin d'obtenir une ligne directrice d'annotation valide et une annotation de référence.

Pour mener leur recherche, les chercheurs ont réuni 30 annotateurs en fonction de critères de base tels que la diversité religieuse, ethnique et résidentielle, afin d'éviter les biais. De plus, les annotateurs devaient remplir les critères suivants :

- 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.

D’après le résultats de cette compagnie d’annotation, voici sont les résultats :

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

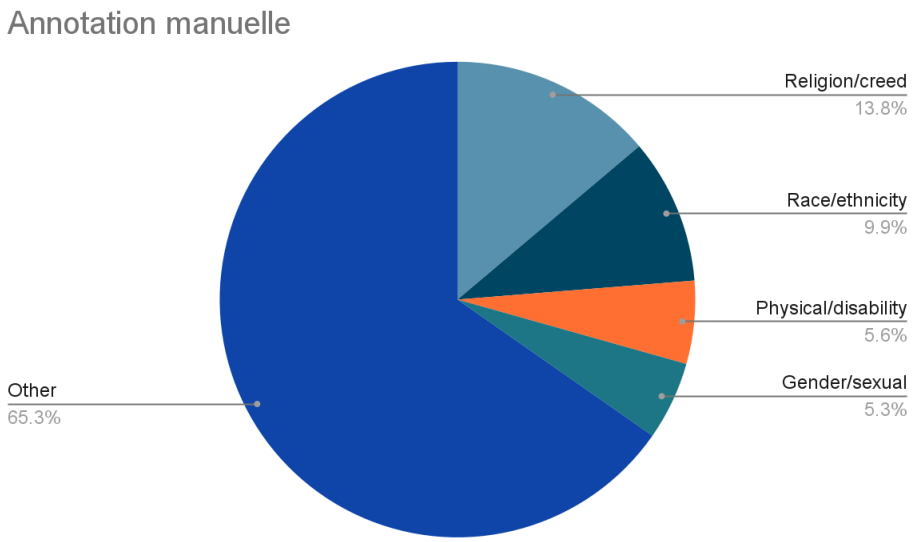

Figure 2 : la répartition des catégories de la haine

Chaque tweet a été annoté par 3 annotateurs pour assurer la meilleure qualité d'annotation.

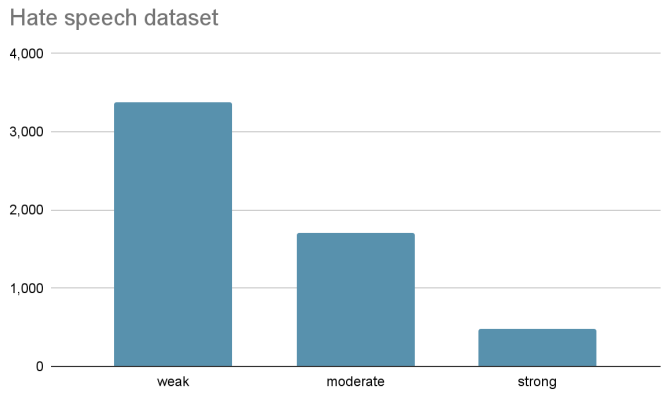

Figure 3 : La répartition des niveaux de la haine

En ce qui concerne les étiquettes de niveau de discours haineux, l'ensemble de données sur les discours haineux se compose de 3 383 discours haineux faibles, 1 705 discours haineux modérés et 473 discours haineux forts.

- **Experiments and Discussions**

L'expérience a 2 scénarios. Le premier, c'est la classification multilabel pour 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 multilabel 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.

En général, le premier scénario et le second scénario ont le même flux, comme le montre le schéma suivant :

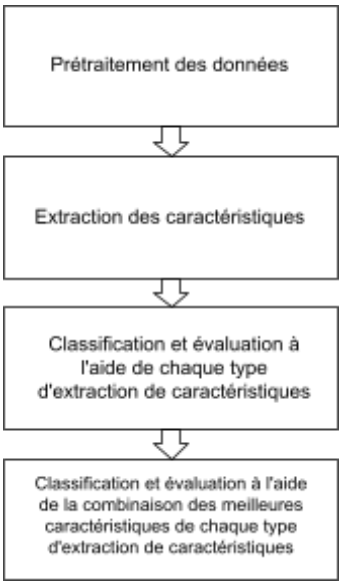

Figure 4 : La chaîne de traitement

1) Le prétraitement des données afin de rendre le processus de classification plus efficace et d'obtenir de meilleurs résultats. Les auteurs ont effectué cinq processus de prétraitement des données, à savoir le pliage des cas, le nettoyage des données, la normalisation du texte (qui transforme les mots non formels en mots formels), le stemming et la suppression des mots vides.

2) Dans cette recherche, les auteurs ont utilisé plusieurs types d'extraction de caractéristiques, à savoir la fréquence des termes (*word n-grams*), l'orthographe (le nombre de points d'exclamation, de points d'interrogation, de majuscules et de minuscules) et le lexique (lexique de sentiment négatif et positif et lexique abusif que les auteurs ont construit eux-mêmes, compilé à partir de mots abusifs.)

3) Pour le classificateur, les auteurs ont utilisé trois algorithmes de classification par apprentissage automatique : Naive Bayes (NB), Support Vector Machine (SVM) et Random Forest Decision Tree (RFDT). Selon des travaux précédents, ces trois algorithmes ont montré de bonnes performances dans la détection des discours haineux et du langage injurieux en indonésien. Cependant, ces classificateurs ne peuvent pas gérer directement la classification de texte en plusieurs catégories. Pour résoudre ce problème, une méthode de transformation des données a été appliquée, permettant aux classificateurs de gérer la classification multicatégorielle. Trois méthodes de transformation de données ont été utilisées : Binary Relevance (BR), Label Power-set (LP) et Classifier Chains (CC). Les évaluations ont été réalisées en

utilisant une technique de validation croisée à 10 volets avec l'exactitude comme métrique d'évaluation.

$$Accuracy = \left(\frac{1}{D} \sum_{i=1}^{D} \left| \frac{\hat{L}^{(i)} \wedge L^{(i)}}{\hat{L}^{(i)} \vee L^{(i)}} \right| \right) \times 100\%$$

Figure 5 : La formule de l'exactitude utilisée par les auteurs de l'étude

1) First Scenario Experiment Result

Les chercheurs ont conclu que les caractéristiques des unigrammes offrent la meilleure précision dans la classification des données, car elles représentent les traits distinctifs de chaque catégorie. Ils ont remarqué que chaque catégorie de classification comporte des mots caractéristiques. Par exemple, dans la catégorie du discours haineux, les tweets étiquetés comme tels contiennent des termes injurieux qui dénigrent un individu ou un groupe, des mots liés à la politique en Indonésie et des mots menaçants/provocateurs.

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	négative sentiment	44.45

Tableau 1 : Meilleure performance de chaque type d'extraction de caractéristiques sur la base de la précision moyenne pour le premier scénario

De plus, en ce qui concerne l'analyse des classificateurs, la méthode d'ensemble sur RFDT s'est révélée plus précise que les méthodes NB et SVM. En ce qui concerne les méthodes de transformation des données, la méthode LP a montré la meilleure précision, car elle permet une corrélation entre les étiquettes uniques, réduisant ainsi les erreurs de classification.

2) Second Scenario Experiment Result

Les chercheurs ont conclu que, lors de l'expérimentation avec différents types d'extraction de caractéristiques, les unigrammes de mots se sont révélés être les meilleures caractéristiques parmi les n-grammes de mots, tandis que les quadrigrammes de caractères

étaient les meilleures caractéristiques parmi les quadrigrammes de caractères. De plus, le point d'exclamation s'est avéré être la meilleure caractéristique orthographique, et la combinaison du sentiment positif et du lexique abusif était la meilleure caractéristique lexicale, selon le deuxième scénario d'expérience. Si l'on examine ces caractéristiques individuellement, le classificateur RFDT avec la méthode de transformation de données LP, en utilisant les unigrammes de mots, a obtenu les meilleures performances avec une précision de 76,16 %.

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

En outre, les chercheurs ont mené une expérience en combinant les meilleures caractéristiques de chaque type. Selon cette expérience, l'utilisation de cette combinaison de meilleures caractéristiques dans le deuxième scénario d'expérience a montré des performances légèrement supérieures par rapport à l'absence de combinaison des meilleures caractéristiques. Le classificateur RFDT avec la méthode de transformation de données LP, en utilisant la combinaison de caractéristiques telles que les unigrammes de mots, les quadrigrammes de caractères, le sentiment positif et le lexique abusif, a offert les meilleures performances avec une précision de 77,36 %.

3) Conclusion

D'après les résultats des deux scénarios d'expérience, les chercheurs ont conclu que la combinaison des unigrammes de mots, du classificateur RFDT et de la méthode de transformation de données LP est la plus performante pour les deux scénarios. Dans le second scénario, leur approche parvient à obtenir une précision de 77,36 % pour la classification de texte à étiquettes multiples visant à identifier le langage abusif et les discours haineux, sans identifier la cible, les catégories et le niveau de discours haineux. Cependant, dans le premier scénario, où l'objectif est d'identifier le langage abusif et les discours haineux, y compris leur cible, leurs catégories et leur niveau, la meilleure performance reste insuffisante avec une précision de seulement 66,12 %. Les erreurs les plus

fréquentes sont des faux négatifs, souvent dues à un déséquilibre dans les données. Les chercheurs envisagent des méthodes d'équilibrage de l'ensemble de données, telles que la collecte de nouvelles données ou le sous-échantillonnage, tout en explorant des approches de classification hiérarchique pour résoudre ce problème complexe de classification à étiquettes multiples. Une autre amélioration serait d'ajouter des caractéristiques sémantiques, comme word embeddings, au processus d'extraction de caractéristiques pour améliorer la précision de cette recherche.

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

- **Introduction**

Dans cette section, nous présentons les résultats de notre expérience, où nous avons appliqué le modèle linguistique développé par les chercheurs à un jeu de données en russe. Nous avons conservé les mêmes définitions pour le langage haineux et abusif, ainsi que les mêmes règles d'annotation.

Pour reproduire le modèle, nous avons utilisé un ensemble de données open source publié sur le site *Hugging Face*[3], composé de 193 tweets en russe susceptibles de contenir des éléments de discours haineux ou de langage abusif. Nous avons ensuite procédé à l'annotation manuelle de chaque tweet en suivant les règles établies dans l'article. À savoir, nous avons d'abord décidé si un tweet peut être classé comme abusif, haineux ou les deux en même temps. Ensuite, pour les tweets haineux nous avons identifié la cible, la catégorie et le niveau.

- **Évaluation de l'annotation**

L'évaluation de l'annotation est essentielle pour mesurer la fiabilité de notre modèle de détection de discours haineux et de langage abusif. Pour cela, nous avons utilisé la mesure de Kappa, qui nous permet d'évaluer l'accord entre les annotateurs. Dans le cadre de notre évaluation, nous avons annoté un échantillon de 50 phrases, chacune étant annotée par deux annotateurs. Cette procédure nous a permis de construire une matrice de confusion pour évaluer la concordance entre les annotations.

[3] https://huggingface.co/apanc/russian-inappropriate-messages

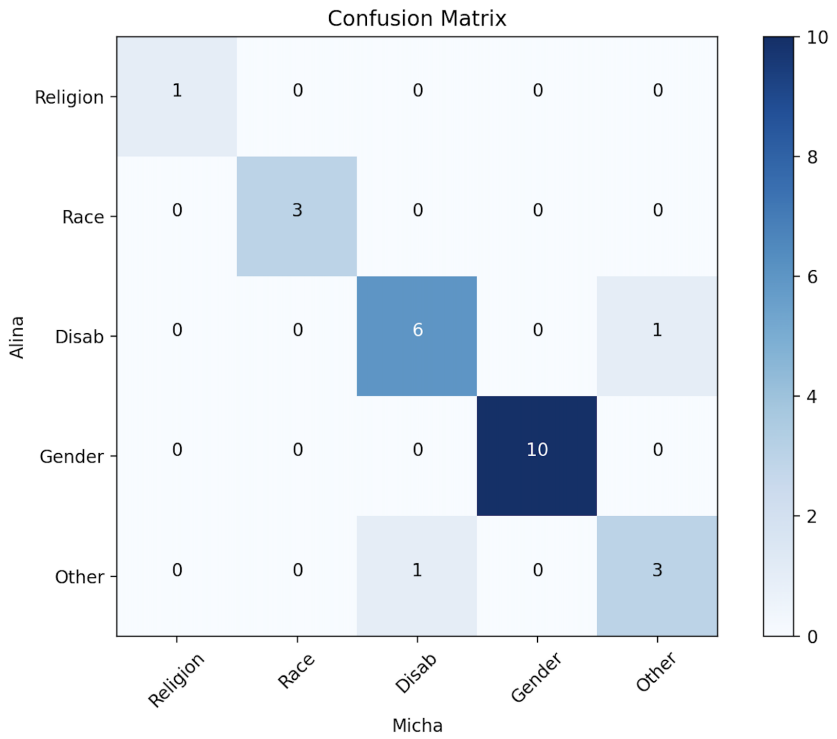

Figure 6 : Matrice de confusion

La figure 6 ci-dessous illustre la matrice de confusion utilisée dans le calcul de Kappa. En utilisant cette matrice de confusion, nous avons calculé les composantes nécessaires pour déterminer le coefficient de Kappa (K) :

Po (Concordant Observations). Le nombre d'annotations concordantes entre les annotateurs. Dans notre cas, il est de 23.

Po =(Concordant Observations) / (Total Annotations)
Po = 23 / 25 = 0,92

Pe (Expected Agreement) : L'accord attendu par le hasard, basé sur les proportions marginales des annotations.

*Pe = (Sum of Your Marginal Proportions) * (Sum of Colleague's Marginal Proportions)*
Pe = 0,00008602 x 2 = 0,00017203

Kappa (K) : Le coefficient de Kappa, qui mesure l'accord corrigé par le hasard. En appliquant la formule, nous obtenons :

Kappa (K) = (Po - Pe) / (1 - Pe)
(0.92 - 0,00017203) / (1 - 0,00017203) = 0,91998624

Un coefficient de Kappa de 0,92 indique un excellent accord entre les annotateurs pour la tâche d'annotation, ce qui renforce la fiabilité de notre modèle. Cela signifie que notre modèle est capable de détecter le discours haineux et le langage abusif avec une précision élevée, ce qui est essentiel pour son utilisation dans des applications de modération en ligne visant à garantir un environnement numérique plus sûr et respectueux.

- **La pertinence des expériences d'apprentissage automatique**

Dans le cadre de notre projet, nous avons entrepris une expérience en machine learning visant à adapter le modèle linguistique élaboré par les auteurs de l'article original pour la détection de discours haineux et de langage abusif. Contrairement à l'approche des auteurs, qui se sont confrontés à deux scénarios distincts, nous avons choisi de nous concentrer sur le second scénario, à savoir la classification multilabel 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. Cette décision découle du constat que même les auteurs n'ont pas pu réaliser le premier scénario de manière satisfaisante, en raison de la complexité de la tâche et du déséquilibre des données.

Pour ce faire, nous avons utilisé un classificateur Support Vector Machine (SVM) afin de déterminer si un tweet contient du langage haineux ou non. Les résultats obtenus sont les suivants :

```
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
```

Table 4 : Résultats SVM

En conclusion, notre expérience en machine learning nous a permis de mettre en évidence les défis liés à la détection de discours haineux, même en se concentrant sur un scénario moins complexe. Les résultats obtenus nous encouragent à explorer davantage de techniques et de modèles pour améliorer la précision de la détection.

- **Analyse**

Une fois l'annotation terminée, nous avons reproduit les calculs des chercheurs pour obtenir le nombre total des tweets haineux, abusifs et haineux et abusifs sans marqueurs de discours de haine. Voici les résultats :

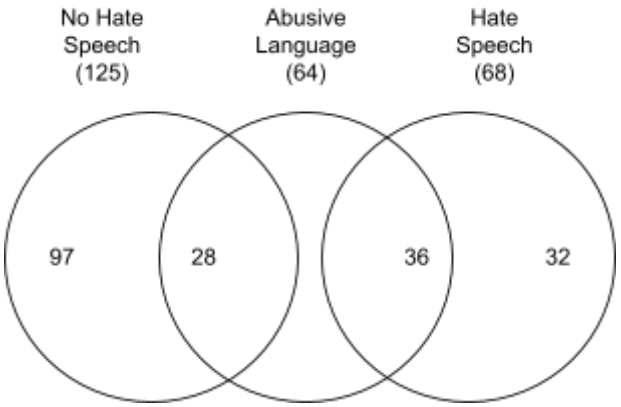

Figure 7 : Répartition du langage abusif entre les tweets ne contenant pas de propos haineux et les tweets contenant des propos haineux

Bien que notre dataset soit largement inférieur à celui utilisé par les chercheurs, nous pouvons observer les mêmes tendances. La plupart des tweets ne contiennent pas les marqueurs de haine, et il y a presque autant de messages abusifs que de ceux haineux.

La prochaine étape consiste à assigner une classe à chaque tweet. Cette partie a suscité plusieurs questions de notre part. L'article indique qu'il est possible d'attribuer plusieurs catégories à un tweet, sauf s'il est classé comme *autre*. C'est-à-dire, par exemple, un message peut se trouver à la fois dans les catégories *religion* et *genre*, mais dans ce cas-là il ne peut pas être dans *autre*. Une question se pose alors si l'on détecte que le même tweet contient les marqueurs des catégories existantes mais aussi d'autres nuances qui ne correspondent pas aux classes définies. Que fait-on ? Est-ce qu'on est censé ignorer ces nuances et plutôt prioriser les catégories préétablies ? En outre, nous n'avons pas trouvé les justifications pour un tel choix de labels. Finalement, l'article montre que le plus grand nombre de tweets ont été classés comme appartenant à la catégorie *autre* (plus de 65%). En même temps, après avoir examiné brièvement à l'aide d'un traducteur le jeu de données des chercheurs nous avons constaté que plusieurs tweets concernent la politique. Vu la prédominance du label *autre*, il serait peut-être utile de revoir les classes pour en choisir les plus adaptées. En ce qui concerne l'application du modèle au dataset russe, étonnamment, le modèle s'avère un peu plus adéquat, même si la catégorie *autre* reste volumineuse. Dans

ce contexte, notre proposition serait d'introduire plus de classes pour une meilleure précision ou d'en remplacer certaines dans les cas où celles-ci sont peu utilisées.

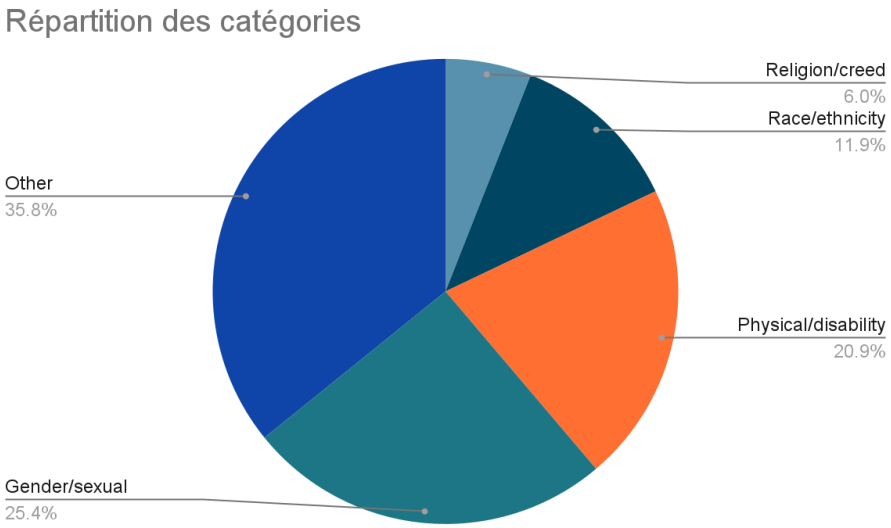

Figure 8 : la répartition des catégories de la haine

Quant aux cibles, dans le jeu de données des auteurs de l'article, la majorité des messages était adressée aux individus, ce qui est le contraire de notre dataset. Il convient de noter cependant, qu'il n'était pas toujours facile de distinguer la cible, car certains textes ciblent à la fois des groupes et des individus, tandis que d'autres restent ambigus et nécessitent plus de contexte.

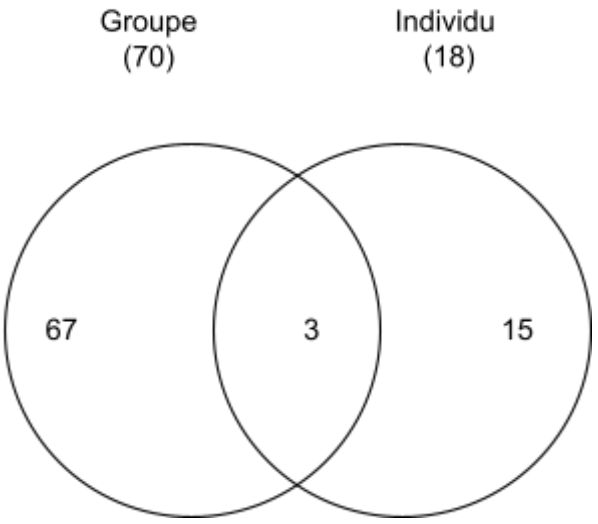

Figure 9 : Répartition de cible de la haine

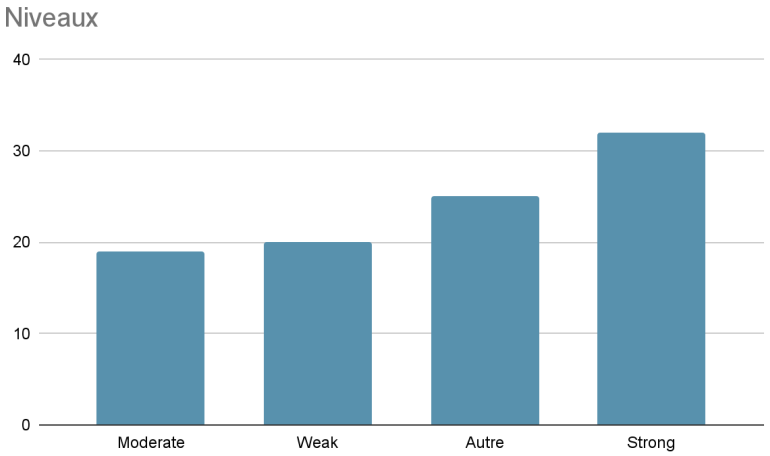

Figure 10 : Répartition des niveaux

3) Conclusion de l'expérience

Notre expérience visait à mettre en pratique le modèle linguistique élaboré par les chercheurs dans l'article initial pour la détection de discours haineux et de langage abusif. Nous avons adapté ce modèle à un jeu de données en russe, en conservant les mêmes définitions pour le langage haineux et abusif, ainsi que les règles d'annotation.

L'analyse des résultats a révélé plusieurs tendances similaires à celles observées dans l'article initial, malgré la différence de taille de nos ensembles de données. La plupart des tweets ne contiennent pas de marqueurs de haine, et il y a une quantité significative de messages abusifs. Cependant, nous avons soulevé des questions concernant l'attribution de catégories aux tweets, en particulier en ce qui concerne les nuances qui ne correspondent pas aux classes définies. De plus, la prédominance de la catégorie "autre" dans le jeu de données initial nous a incités à reconsidérer les classes utilisées, en particulier lorsque celles-ci sont peu utilisées. En autre, on constate également la différence de notre expérience avec une majorité de messages adressés à des groupes plutôt qu'à des individus.

En conclusion, cette expérience a permis de montrer la robustesse du modèle de détection de discours haineux et abusif, même lorsqu'il est appliqué à une langue et un contexte culturel différents. Cependant, des ajustements peuvent être nécessaires pour mieux correspondre à la réalité du langage russe en ligne. Il est essentiel de continuer à améliorer et à affiner ce modèle pour garantir une détection plus précise et une meilleure classification des messages, en vue de contribuer à la création d'un environnement numérique plus sûr et respectueux pour tous les utilisateurs de la langue russe.