Cyberharcèlement et discours de haine : construction, annotation et analyse d'un corpus extrait des réseaux sociaux

Mémoire de Master Lettres Parcours Linguistique, traitement informatique du texte et processus cognitifs

Soutenu par **Manon Audren**

Le 16 juillet 2021

Dirigé par Mme Elena Cabrio

Jury : Mme Blaya, Mme Cabrio, Mme Ollagnier, Mme Passino et Mme Villata

Remerciements

Je tiens à remercier l'Université de Nice Côte d'Azur pour m'avoir permis de faire mon stage et notamment Mme Diana Passino, la responsable du Master, pour m'avoir soutenue, permis de trouver mon stage rapidement et d'avoir toujours pris le temps de répondre à mes questions.

Je remercie tout particulièrement Mme Cabrio, enseignant-chercheur à l'Université Côte d'Azur et au laboratoire I3S de Sophia Antipolis pour avoir accepté de diriger mon mémoire, pour sa disponibilité lorsque j'en avais besoin, pour m'avoir de nouveau acceptée en tant que stagiaire ainsi que pour m'avoir partagé ses connaissances en linguistique informatique afin de les mettre en pratique.

Je souhaite également remercier l'équipe Wimmics pour m'avoir gentiment accueillie pendant ces quelques mois au sein de leur équipe, toujours avec bienveillance et écoute.

Enfin, je remercie ma famille et mes amies pour leur soutien et leur relecture.

Résumé

Le cyber-harcèlement est de plus en plus fréquent en ligne et concerne également plus de personnes. Afin de repérer ces messages de haine sur les réseaux sociaux et sanctionner les auteurs, les chercheurs en traitement automatique du langage tentent d'appliquer leur algorithme à ces recherches. Seulement, des données sont nécessaires pour entraîner ces algorithmes, c'est pourquoi notre étude consiste à en recueillir un échantillon en recueillant des données sur Twitter et lors d'une expérimentation avec des élèves, annoter les messages (cibles, sujets de hate speech, présence ou non d'humour et de messages implicites pour les messages de l'expérimentation) et analyser ces données pour constituer un corpus et former l'algorithme.

Les analyses montrent des stratégies différentes pour exprimer de la haine envers un sujet, que ce soit, l'insulte directe, l'humour ou les insultes implicites comme les métaphores ou les sous-entendus.

A partir de ces résultats, les chercheurs vont donc pouvoir former leur algorithme et ainsi aider les plateformes en ligne à repérer et réduire le discours de haine.

Cyber-harassment is on the rise online and is affecting more people as well. In order to spot these hate messages on social media and punish the perpetrators, automatic language processing researchers are trying to apply their algorithm to this research.

But, data is needed to train these algorithms, that's why our study consists of collecting a sample by collecting data on Twitter and during an experiment with students, annotate the messages (targets, subjects of hate speech, presence or not of humor and implicit messages for the messages of the experiment) and analyze these data to constitute a corpus and train the algorithm.

The analyzes show different strategies for expressing hatred towards a subject, whether it is direct insult, humor or indirect insults such as metaphors or innuendo. From these results, the researchers will therefore be able to train their algorithm and thus help online platforms to identify and reduce hate speech.

Table des matières

Introduction

Aujourd'hui, à l'heure des nouvelles technologies et de l'expansion des réseaux sociaux, nous nous trouvons de plus en plus exposés en ligne. En effet, selon un rapport de We Are Social / Hootsuite publié en février 2021, 49,6 millions de français sont des utilisateurs actifs des médias sociaux, soit 75,9% de la population. Les Français passent en moyenne 5h37 par jour à utiliser internet, dont 1h41 à utiliser les réseaux sociaux. De plus, les utilisateurs des médias sociaux possèdent en moyenne 6,8 comptes. Le fait d'être exposé en ligne induit plus de "chances" de se faire harceler, que ce soit par mails, sur les réseaux sociaux… Ce type de harcèlement est appelé "cyber-harcèlement" et il s'agit d'un délit. Le site du service public nous indique que si nous sommes victime de ce type de harcèlement, nous pouvons demander le retrait des publications à leur auteur ou au responsable du support technique. Malheureusement, il existe des plateformes où le nombre de messages contenant du cyber-harcèlement est trop important pour être contrôlé ou alors parce que la plateforme ne possède pas de système de contrôle pour ce type de harcèlement. C'est le cas de Twitter, qui, étant donné le nombre considérable de nouveaux tweets par jour (environ 500 millions selon Chiffres Twitter - 2021), ne peut pas contrôler le contenu de tous les tweets, c'est pourquoi, dans les règles du site il est mentionné que tout tweet offensant doit être signalé pour qu'un modérateur l'analyse, le supprime ou suspende le compte de l'auteur si besoin.

Notre objectif est d'identifier et prévenir les éventuels impacts négatifs du cyberharcèlement sur les jeunes, en mettant au point des technologies de pointe pour la détection précoce de ces phénomènes grâce au suivi des médias sociaux. Pour ce faire, nous allons utiliser le Traitement Automatique du Langage (TAL), qui est une branche de l'Intelligence Artificielle (IA) et qui permet de créer des outils qui vont traiter le langage naturel pour divers usages. Dans les dernières années, les chercheurs en TAL ont montré un grand intérêt dans l'application de leur algorithme dans le but de détecter les messages de haine. Il s'agit d'algorithmes supervisés, qui vont apprendre à prédire quelque chose à partir d'exemples annotés. Ainsi, les exemples annotés constituent la base d'apprentissage de l'algorithme, et il va ensuite devoir généraliser cet apprentissage à d'autres unités.

Notre travail consiste donc à “aider” cet algorithme en trouvant et annotant des exemples sur lesquels il va se baser, pour ensuite repérer tout seul les messages de haine. Des corpus annotés de haute qualité sont donc nécessaires pour être utilisés comme données d’entraînement par ces outils.

Le but de ce mémoire consiste donc à extraire un échantillon de messages courts échangées en ligne par rapport à des cibles de hate speech définies (que ce soit directement sur les réseaux sociaux ou lors d’une expérimentation avec des jeunes), de détecter si ces messages contiennent bien du hate speech, de les annoter et classifier en fonction de la cible visée et d’extraire ainsi que comprendre la subtilité des structures argumentatives des messages haineux, en allant au-delà de la simple détection d’insultes. Nous tenons à préciser que ces données recueillies ne sont pas exhaustives, que ce soit pour ce qui est des thèmes abordés mais également des personnes visées.

Le mémoire est structuré comme il suit :

Dans un premier temps nous verrons comment nous avons extrait des tweets, qui sont des messages ponctuels et qui visent des groupes en général, pour créer notre propre corpus, l’analyse des données obtenues sur quelles sont les cibles des messages de haine sur Twitter, les principales raisons de ces messages, plusieurs analyses des corrélations entre ces catégories et la présence ou non d’humour.

Dans une deuxième partie, nous vous expliquerons comment nous avons recueilli des messages de haine lors d’une expérimentation, qui sont eux, contrairement aux tweets, des messages plus insistants et personnels qui représentent bien le hate speech, ainsi que l’analyse des personnes visées, si les victimes sont les principales harcelées ou si les harceleurs aussi ont subi du hate speech, nous verrons également quels sont les sujets de hate speech présents parmi les thèmes étudiés, s’il y a également ou non de l’humour dans ces messages mais aussi si les auteurs utilisent des tournures implicites pour s’exprimer.

Enfin, nous conclurons ce mémoire par un rappel des résultats et ce qu’il serait intéressant de faire dans la continuation de ce travail.

Etat de l'art

Plusieurs études ont déjà été faites sur le cyber-harcèlement et les messages de haine dans d'autres langues comme l'anglais, mais pas le français, c'est pourquoi, nous nous penchons sur l'étude de cette langue. Au commencement de notre travail, il est essentiel de voir ce qui a été fait, ce qui ne l'a pas été ainsi que ce qui fonctionne ou pas pour pouvoir réussir notre étude. Ainsi, compte tenu des articles que nous pouvons trouver, nous avons pu voir que certains constituent une approche sociologique et d'autres, une approche computationnelle.

A. Approche sociologique et définitionnelle

Tout d'abord il est important de définir le cyber-harcèlement afin de mieux le repérer dans le discours ainsi que d'évaluer la gravité de ce qu'il représente. La plupart des articles qui le présentent, le définissent pratiquement tous de la même manière, c'est-à-dire comme étant "des propos diffamatoires, du harcèlement ou de la discrimination, la divulgation d'informations personnelles ou des propos humiliants, agressifs, vulgaires" (Cyberviolence et cyberharcèlement: approches sociologiques). Un autre article de Catherine Blaya (Le cyberharcèlement chez les jeunes) en donne une définition plus simple : "le cyberharcèlement fait référence à la violence entre jeunes ou à celle à laquelle ils sont potentiellement soumis lorsqu'ils surfent sur internet" ou le cyberharcèlement correspond à "toute forme de violence en ligne". Mais, dans ce même article elle explique aussi que toutes les définitions ne prennent pas en compte la répétition comme un critère du cyber-harcèlement : "certains estiment qu'en raison de la permanence des messages en ligne et par conséquent d'une exposition qui s'inscrit dans la durée tant pour les victimes que pour les témoins, il est possible de parler de harcèlement, même sans répétition de l'acte, les conséquences pouvant être aussi sévères que pour les victimes répétées". Donc tous les auteurs ne sont pas d'accord sur le fait que la répétition est ou non un critère à prendre en compte pour définir un message comme étant du cyber-harcèlement mais ils sont tous d'accord sur le fait que c'est un message de violence, de haine envers une ou plusieurs personnes.

Mme Blaya nous expose les aspects plus sociologiques du cyber-harcèlement dans ses articles. En effet, comme nous pouvons le voir dans Cyberviolence et cyberharcèlement: approches sociologiques, elle se questionne sur ce qu'est le cyber-harcèlement mais aussi quels peuvent être les profils des victimes et des agresseurs afin de mieux intervenir, prévenir et réduire ce harcèlement. Selon une étude réalisée sur la cyberviolence entre les collégiens en France, "18,4% des participants à la recherche s'estiment victimes de cyberviolence". Cette étude indiquerait aussi que les victimes sur internet le seraient également dans la vie réelle.

Pour ce qui est des causes du cyber-harcèlement, l'article Chapitre 2. Le discours de haine explique que ce sont des "stéréotypes négatifs qui appréhendent certains groupes ou individus comme inférieurs, différents et moins dignes de respect".

Tout cela nous montre bien que le cyber-harcèlement ou le harcèlement sont des violences qui touchent de nombreux jeunes français dans la vie de tous les jours et qu'ils ont besoin de l'algorithme que nous souhaitons développer pour réduire le plus possible ce qu'ils vivent sur internet, mais aussi que nous sensibilisions davantage les générations à venir pour réduire le plus possible ce fléau pouvant avoir un impact fort dans la vie de quelqu'un. Un récent article publié par IFOP (Harcèlement entre pairs en milieu scolaire, quelle est l'ampleur de ce phénomène ?), montre que "63% des personnes ayant été harcelées pendant plus de deux ans en gardent des séquelles psychologiques". De plus, 95% des sondés estiment que ce phénomène est en augmentation et 92% jugent qu'il n'est pas appréhendé à sa juste valeur par les pouvoirs publics en France, ce qui montre bien que notre projet est important pour qu'un corpus de données en français soit disponible, mais aussi pour réduire ce harcèlement et peut-être, ainsi améliorer la vie de ces personnes.

B. Approche computationnelle

D'autres articles traitant du cyber-harcèlement, ont eux, comme nous l'avons dit précédemment, une approche plus computationnelle. En effet, les experts en TAL sont ravis de pouvoir contribuer à ces recherches actuelles et de pouvoir aider avec leurs systèmes, à réduire ce type de harcèlement, puisque comme nous pouvons le voir dans l'article Chapitre 3. Discours de haine en ligne et réseaux sociaux, qui traite de la réglementation du discours de haine sur les réseaux sociaux, nous apprenons que "signaler et retirer le discours de haine constitue une stratégie cruciale pour le combattre ; mais elle ne suffit pas". Effectivement, il y a des formes de discours de haine qui n'entrent pas dans une catégorie, celles liées au contexte, l'humour… et donc ne peuvent être retirées. De plus, les auteurs de cet article nous expliquent que maintenant, les auteurs des discours de haine "connaissent les règles des réseaux sociaux et leurs failles et ont mis au point des tactiques élaborées pour faire passer leurs messages". Puisque le discours de haine se fait de plus en plus discret vis à vis des règles de réglementation des réseaux sociaux, les experts en TAL se doivent de suivre ces aspects afin de toujours améliorer leurs systèmes pour que le discours de haine soit repérable au maximum.

Les auteurs de l'article Mik3M4n/help_dicrah: A ML-based hate speech detection tool on Twitter (in French) ont réussi à créer un outil en ligne pour détecter les discours de haine lors d'un flux Twitter, seulement, étant donné de grosses contraintes de temps, ils n'ont pas pu relever assez de tweets pour former au mieux l'algorithme, ni annoter au maximum chacun des tweets. Nous ne pouvions donc pas utiliser leurs données, c'est pourquoi, nous avons dû nous même constituer un corpus entier.

Nous pouvons voir dans l'article A Multilingual Evaluation for Online Hate Speech Detection qu'une nouvelle manière de repérer le langage abusif est utilisée et semble fonctionner "de manière satisfaisante dans différentes langues comme l'Anglais, l'Italien et l'Allemand". Il s'agit de l'utilisation d'une architecture neuronale ainsi que des composants qui prennent en compte ce qui peut s'ajouter au texte comme les émoticônes, les émotions, les hashtags…

Mais ce qui pose encore problème, c'est la catégorisation des messages de haine. C'est pourquoi le projet intitulé "L'intelligence artificielle au service de la prévention de la cyberviolence, du cyber-harcèlement et de la haine en ligne" (OTESIA) de l'appel UCA IDEX OTESIA (dans la suite du projet CREEP) (Creep: Home & DELIVERABLE CREEP2), pour lequel nous travaillons, a aussi été créé, pour aller jusqu'au bout dans le développement de l'algorithme supervisé permettant de repérer le hate speech, de le classifier à grains fins pour identifier au mieux les communautés ciblées mais également, pour rendre disponible un corpus de données en langue française, puisque jusqu'ici, seuls des corpus de hate speech dans d'autres langues sont disponibles.

Comme point de départ, les travaux actuels dans le cadre de ce projet se basent sur un corpus comprenant des données de tweets disponibles gratuitement (HKUST-KnowComp/MLMA_hate_speech: Dataset and code of our EMNLP 2019 paper "Multilingual and Multi-Aspect Hate Speech Analysis") mais, certains thèmes ne sont pas assez représentés pour pouvoir au mieux identifier les communautés, c'est pour cela que notre travail est utile: il va permettre de récolter de nouvelles données (sur Twitter et grâce aux messages échangés entre les élèves de l'expérimentation) pour compléter les thèmes ayant des données insuffisantes et ainsi contribuer à l'amélioration d'un algorithme supervisé qui repère le hate speech et les cibles de celui-ci. Notre expérimentation en école nous a également permis de faire de la prévention sur le cyber-harcèlement et ses effets sur les élèves, grâce à la présence d'une sociologue.

I. Partie 1 : Catégorisation de tweets à grains fins et analyse

A. Méthode de recueil des données

Pour pouvoir réaliser notre étude sur le cyberharcèlement et les messages de haine, nous avons utilisé un jeu de données existant mais qui ne comportait pas assez de tweets pour les catégories Gender, Religion et Sexual Orientation :

HKUST-KnowComp/MLMA_hate_speech: Dataset and code of our EMNLP 2019 paper "Multilingual and Multi-Aspect Hate Speech Analysis". Nous avons donc dû extraire nous même des tweets pour compléter ces catégories et c'est pour cela que nos données recueillies ne sont pas exhaustives. Tout d'abord, nous avons récupéré un ensemble de plus de 3000 tweets sur le thème "Gender" (extraits de An Annotated Corpus for Sexism Detection in French Tweets), dans lequel nous avons supposé qu'il pouvait également y avoir des messages de haine, les repérer et les relever pour commencer à créer un corpus avec uniquement des tweets ayant des messages haineux.

Ensuite, nous avons composé une liste de hashtags et de mots-clés (dont nous pouvons voir un extrait ci-dessous) en rapport avec les thèmes que nous devions traiter (c'est à dire, Gender, Religion et Sexual Orientation), qui nous permettront d'extraire des tweets les contenant grâce à l'outil de recherche Twitter.

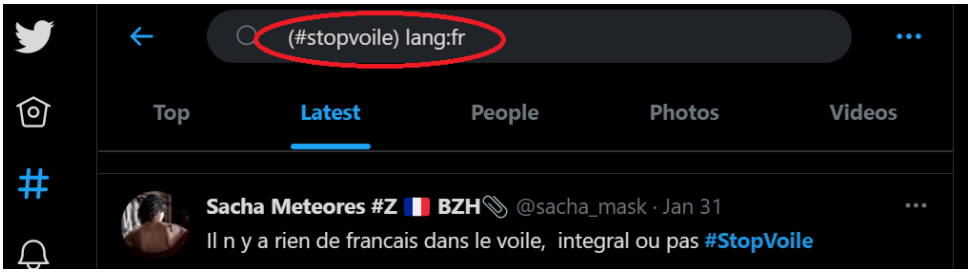
Mots-clés Gender	Mots-clés Religion	Mots-clés Sexual Orientation
gonzesse.s	religion haineuse	sida
salop(s).e(s)	islamophobie	encule.r/enculé.s
pute.s	cathophobe.s	sodomie.s
connasse.s	décapite.r	con.s/conne.s
connard.s	profane.r/profanation.s/profané(s).e(s)	fiotte.s
con(s).ne(s)	immigration	pédale.s
enceinte	migrant.s	orgie.s
consentement	christianophobe.s	capote.s
tenue	fachos	homophobe.s
viol.er/violé.e	racisme	pd
attouchement.s	raciste.s	fdp
	idéologie	LGBT
	voilé(s).e(s)	gays
	dhimmitude	suceur.se
		sous race
		fils de pute
		coming out
		feminisation

Liste des mots-clés utilisés pour trouver des tweets

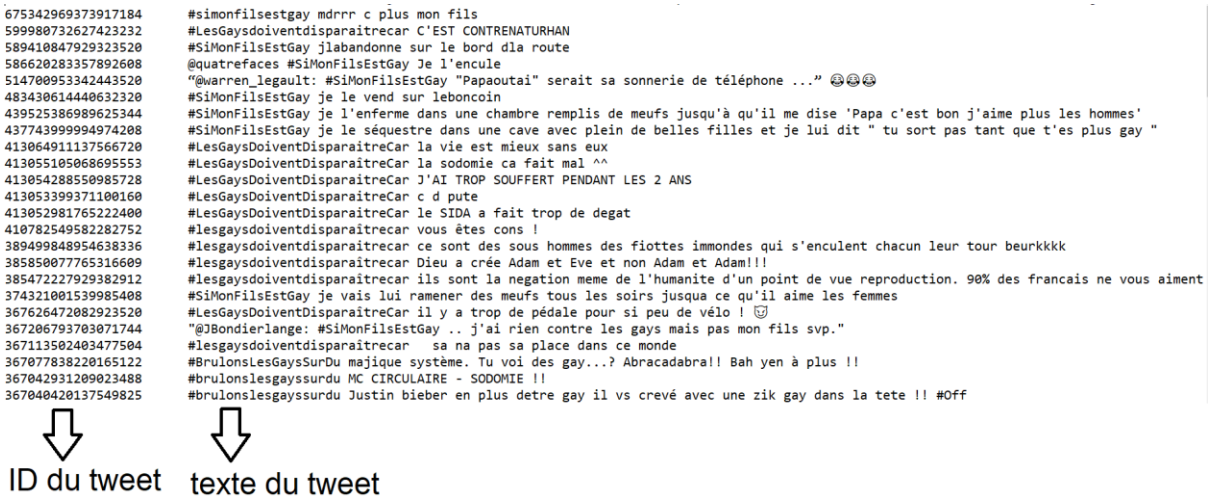
#Gender	#Religion	#SexualOrientation
#sexiste	#unbonjuif	brulonslesgayssurdu
#leshommessontdessalauds	#agoodjew	brulonslesgays
#slt	#unjuifmort	lesgaysdoiventdisparaitrecar
#balancetonporc	#christianophobie	simonfilsestgay
#payetashnek	#stopislam	tesunpd
#moiaussi	#stopimmigration	simafilleestlesbienne
#metoo	#remigration	antigay
#stopharcelementderue	#grandremplacement	#LGBT
	#migrants	#simonfilsestgay
	#soutiengénérationidentitaire	#lesgaysdoiventdisparaitrecar
	#islamogauchisme	#brulonslesgayssurdu
	#stopauvoile	#ungaymort
	#stopterrorisme	#antilesbienne
	#islamconquérent	#antigay
	#interdictionduvoile	#tantouze
	#dérivesectaire	#grossetapette
	#mafrancepaslatienne	#sodomie
	#antiislam	#simafilleestlesbienne
	#contrelislam	#brulonslesgays
	#islamophobe	#lesgaysdoiventdisparaitre
	#antijudaïsme	#lesbienne
	#guerredereligion	#tafiole
	#antisémitisme	#pédale
	#antichristianisme	#saloperie

Extrait de la liste des hashtags utilisés pour trouver les tweets

Cette liste nous a ensuite permis de trouver de nouveaux tweets directement sur Twitter, avec l'outil de recherche disponible. Nos conditions pour récolter un tweet étaient qu'il devait être en français et comporter du hate speech envers un des trois thèmes. Pour se faciliter la tâche et donc valider ces deux conditions pour chaque recherche, on entrait une formule sur Twitter, comme ci-dessous, et il nous suffisait de remplacer le terme entre parenthèses pour obtenir des tweets sur un thème différent ou comportant un hashtag différent. Par exemple #stopislam si nous souhaitions un message sur la religion ou #balancetonporc si nous souhaitions plutôt un message sur le genre.



Lorsqu'un tweet validait ces conditions, on copiait l'identifiant et le texte de celui-ci dans un document, comme nous pouvons le voir dans l'image ci-dessous, qui est un extrait du corpus Sexual Orientation.



Finalement, nous avons trois documents textes (un pour chaque thème) comportant 200 messages de haine chacun, ce qui constitue notre corpus final.

B. Analyse des données recueillies

1. <u>Analyse des données "cibles"</u>

Dans cette partie, nous allons voir selon le thème des tweets, quels sont les profils visés ainsi que ceux qui le sont plus ou moins, grâce à une quantification en pourcentages.

Ainsi, chaque thème (gender, religion et sexual orientation) sur lequel nous avons travaillé comporte un corpus de 200 tweets hate speech. Sur ces 200 tweets, on a observé quelles sont les cibles du hate speech et comptabilisé leurs récurrences, c'est-à-dire, le nombre de tweets qui visaient telle ou telle cible. Enfin, grâce aux récurrences, nous avons pu calculer une valeur en pourcentages pour que les chiffres soient plus parlants et puissent au mieux rendre compte des observations. Nous avons donc multiplié la valeur de récurrence par cent et divisé le tout par le nombre total de tweets du document en question (**exemple** : pour le document HS GENDER : (4 (= nombre de tweets ayant les hommes pour cible) x 100) / 200 (= le nombre total de tweets de ce document) = 2%).

GENDER

Dans ce corpus de tweets, il y a beaucoup de tweets comportant des hashtags que ce soit #balancetonporc ou #payetashnek par exemple. #payetashnek est explicite, nous savons que les personnes utilisant cet hashtag en tant que victimes sont des femmes étant donné que le sexe est explicite dans celui-ci (shnek = sexe féminin). Mais pour #balancetonporc, le sexe n'est pas explicite, donc il peut aussi bien s'agir de femmes mais aussi d'hommes qui désignent une agression par un homme. Il se peut aussi que certains utilisent cet hashtag pour désigner une agression faite par une femme, si la victime veut que son tweet soit plus vu, car même si l'hashtag #balancetatruie existe, il est peu utilisé et donc peu référencé par rapport à #balancetonporc.

Ainsi pour le thème "gender", les cibles sont les suivantes et nous nous sommes inspirés des catégories utilisées dans un article sur l'identification automatique de la misogynie :

- "Hommes" : pour les tweets visant explicitement et uniquement les hommes.

Exemple : "Tous les mêmes! #leshommessontdessalops #Facebook #drague #payetashnek" les hommes sont ici explicitement visés à travers l'hashtag et l'insulte "#leshommessontdessalops".

- "Femmes" : pour les tweets visant explicitement et uniquement les femmes.

Exemple : “Vous vous dites femmes vous savez même pas faites des pâtes 🍝 bande de connasse” ici la femme en général est explicitement insultée de “connasse” et est considérée comme incapable de faire à manger.

- “les deux” : pour les tweets visant explicitement les hommes et les femmes.

Exemple : “Morata va t'occuper de ta putain femme mais arrête de jouer au foot” dans ce tweet on peut voir du hate speech explicite envers les deux sexes : la femme que l’on insulte de “putain” mais aussi l’homme de manière plus subtile où on devine que l’auteur du tweet l’envoie promener en disant ”va t’occuper de…” et, on comprend bien qu’il faut que “Morata” arrête de jouer au foot, sûrement parce qu’il n’est pas doué.

- “Cible non définie” : pour les tweets qui visent quelqu’un mais on ne sait pas si c’est un homme, une femme ou si ce tweet vise les deux sexes. Pour beaucoup de tweets de cette catégorie, on suppose que les cibles sont des femmes mais nous ne sommes pas sûrs à 100%. De plus, il se peut aussi que ce soient des hommes.

Exemple : “la police franchement ils abusent quand il me touche le cul au premier rdv #BalanceTonPorc” dans ce tweet, aucun indice ne laisse entendre le sexe de la personne qui l’a rédigé, l’hashtag “#balancetonporc” sous-entend fortement qu’une femme est la cible mais il peut arriver que des hommes l’utilisent également. Nous ne pouvons donc pas déterminer si la victime est un homme ou une femme.

Document	Cibles	Récurrences	Total	Pourcentage
HS Gender	Hommes	4	200	2,0%
	Femmes	174		87,0%
	Les deux	5		2,5%
	Cible non définie	17		8,5%

Pour ce thème, la catégorie la plus visée dans les 200 tweets est de loin celle des femmes avec 87%. Celle qui est la moins visée est “hommes” avec seulement 2%. On remarque aussi que la catégorie des hommes plus les femmes (2,5%) est plus visée (légèrement) que celle des hommes uniquement.

Comme le montrent ces résultats, la catégorie “femmes” est le genre qui subit le plus le sexisme, sur Twitter notamment. En effet, il était beaucoup plus simple de trouver des tweets visant des femmes étant donné les nombreux hashtags récents qui permettent de dénoncer le sexisme ou même des agressions subies par les femmes, alors que pour les hommes, il n’existe pas de hashtags comme tel, soit parce qu’ils n’ont pas été créés et donc que les hommes victimes utilisent les “hashtags féminins”, soit parce que les hommes subissent moins le sexisme que les femmes, et donc, dans ce cas notre étude est représentative.

RELIGION

Les tweets de ce thème comportent des hashtags explicites qui suffisent à catégoriser la cible du tweet (**exemple** : #stopislam, #unjuifmort…). Dès lors qu’un de ces hashtags est présent dans un tweet, il est catégorisé dans la cible correspondante, sauf contre-indication. Mais, si un tweet en comporte plusieurs, il est catégorisé dans “cible non définie”, sauf si le contexte explicite la visée d’une cible plus qu’une autre.

Pour ce thème, les cibles relevées sont :

- “Judaïsme” : pour les tweets visant uniquement cette religion ou ses pratiquants, que ce soit le courant judaïque orthodoxe, le judaïsme conservatif ou Massorti et le judaïsme réformé.

Exemple : “@yhNico: #UnBonJuif enlève le gaz avant de boire du coca”, ce tweet est effectivement du hate speech envers les juifs compte tenu de l’association de l’hashtag comprenant la cible, “#unbonjuif”, et de la suite de mots “enlève le gaz” qui font référence au génocide subit par les juifs entre 1941 et 1945.

- “Islam” : pour les tweets visant uniquement cette religion ou ses pratiquants, que ce soit la branche des Sunnites, des Chiites, des Khâridjites…

Exemple : “L'ISLAMISME VEUT NOTRE PEAU... #StopIslam #StopImmigration” dans ce tweet la cible est on ne peut plus claire entre la présence du mot “ISLAMISME” et de l’hashtag “#StopIslam”.

- “Christianisme” : pour les tweets visant uniquement cette religion ou ses pratiquants, que ce soit la branche des Catholiques, des Protestants ou des Orthodoxes.

Exemple : “⚫ 67 tombes chrétiennes profanées à #Fontainebleau par des tags nazis L’#antichristianisme n’a pas sa place en France ! Au pouvoir dès 2022, le #RN mettra en œuvre des lois sévères pour ceux qui ne respectent pas la civilisation française.”, entre la profanation des tombes et l’hashtag #antichristianisme, la cible est bien explicite.

- “Cible non définie” : pour les tweets visant une ou plusieurs religion(s) mais on ne sait pas laquelle/lesquelles précisément. Les tweets visant plusieurs cibles au même niveau sont également dans cette catégorie.

Exemple : “Sur 10 personnes assassinées chaque jour dans le monde en raison leur foi, 8 sont chrétiennes, 2 sont juives. Dans les 10 cas, les assassins sont musulmans. #Islam extrémiste, #antisemitisme et #antichristianisme explosent. Dans le silence assourdissant des medias bienpensants 😔”, ici nous pouvons voir que les trois religions sont autant visées l’une que l’autre : le christiannisme est visé de par les actes violents et le hashtag #antichristiannisme, le judaïsme est aussi visé par les actes violents et le hashtag #antisémitisme et l’islam est visé avec le hashatg #islam suivi du mot “extrémiste”.

Document	Cibles	Récurrences	Total	Pourcentage
HS Religion	Judaïsme	28	200	14,0%
	Islam	128		64,0%
	Christianisme	35		17,5%
	Cible non définie	9		4,5%

Pour ce thème, la catégorie la plus visée est “Islam” avec 64%. Le Christianisme est visé à 17,5% et le Judaïsme à 14%. Donc, comme nous pouvons le voir, la religion qui subit le plus le racisme sur Twitter est l’Islam.

SEXUAL ORIENTATION

Dans plusieurs tweets de ce thème, il n'y a pas d'indice sur la cible à part l'hashtag ou mot "tantouze", "tapette"… par exemple. Ces hashtags désignent généralement les hommes homosexuels, donc ces tweets ont pour cible les hommes, à moins qu'il y ait un indice comme quoi la cible serait une femme (**exemple** : je suis seulE et j'ai peur #tantouze).

La catégorie "sexual orientation" comporte les mêmes cibles que "gender" :

- "Hommes" : catégorie de tweets qui visent uniquement les hommes homosexuels.

Exemple : "#simonfilsestgay mdrrr c plus mon fils" où la cible est très claire avec le nom "fils" répété deux fois.

- "Femmes" : catégorie de tweets qui visent uniquement les femmes homosexuelles.

Exemple : "#SiMaFilleEstLesbienne , elle est tout simplement dans la merde ahahah !" la cible "fille" est bien distincte dans ce tweet, de plus il y a également l'adjectif "lesbienne" qui désigne une femme homosexuelle.

Dans le cas de tweets où il n'y a pas de mot explicite désignant une personne de sexe féminin ou masculin (**exemple** : femme, fille, meuf, gonzesse, mec, homme…), nous devons chercher des indices : un adjectif au féminin, un participe passé… (**exemple** : grosSE : "se" est la marque du féminin ici) afin de nous aider à déterminer le sexe de la cible visée. En effet, si nous trouvons un indice comme tel, nous pouvons être sûr que l'auteur a voulu désigner une femme puisque cela demande une réflexion, un effort d'accorder l'adjectif avec la personne qu'il désigne (une femme en l'occurrence) ce qui n'est pas le cas pour le masculin. Pour les messages où il n'y a pas de mot explicite désignant le sexe du sujet et s'il n'y a pas non plus d'indice avec les participes passés ou les adjectifs, nous ne pouvons, par contre, pas être sûr que la personne désignée est un individu masculin, puisqu'il peut s'agir d'une femme, mais que l'auteur a oublié d'accorder par exemple. Dans ce cas, le tweet est catégorisé comme "cible non définie".

- "les deux" : qui vise explicitement les hommes homosexuels et les femmes homosexuelles.

Exemple : “#BrulonsLesGaysSurDu lady gaga, qu'ils aillent tout droit vers l'enfer avec leur pédale d' idole” dans celui-ci la cible est plus implicite, on voit bien que les femmes sont visées avec “Lady Gaga” et “pédale d’idole”, pour les hommes on les retrouve dans le déterminant “ils” puisque ce déterminant désigne soit uniquement des hommes ou des femmes plus au moins un homme, on peut donc voir que les deux sexes sont ici visés.

- “Cible non définie” : on ne sait pas explicitement qui est visé. En général le terme “gay” est utilisé, or certains considèrent que cela signifie “homme homosexuel” et d’autres “personne (donc homme ou femme) homosexuelle”. Puisqu’il y a une ambiguïté je les catégorise dans cette catégorie sauf contre-indication.

Exemple : “#lesgaysdoiventdisparaîtrecar sa na pas sa place dans ce monde” comme dit précédemment, le terme “gay” est employé et aucun autre indice ne laisse sous-entendre si c’est un homosexuel qui est visé ou une homosexuelle.

Document	Cibles	Récurrences	Total	Pourcentage
HS Sexual Orientation	Hommes	63	200	31,5%
	Femmes	14		7,0%
	Les deux	4		2,0%
	Cible non définie	119		59,5%

La catégorie la plus visée est “cible non définie” avec 59,5%. Nous ne savons donc pas si ce sont plus les hommes, les femmes ou les deux d’une manière générale qui sont le plus visés, étant donné que la plupart des tweets de la catégorie “cible non définie” comportent le terme “gay” et aucun indice nous permet de savoir si ce terme désigne alors un homme, une femme ou s’il est utilisé pour désigner les deux sexes. Sinon, la deuxième catégorie la plus visée est “hommes” avec 31,5%. Donc, par rapport aux femmes, les hommes homosexuels subissent plus le harcèlement et le hate speech, au moins sur Twitter.

2. Analyse des données “sujets”

Dans cette partie, nous allons voir selon le thème des tweets, quels sont les sujets de hate speech des tweets, ceux qui sont plus fréquents et ceux qui le sont moins, grâce à une quantification en pourcentages, comme précédemment.

Ainsi, sur les 200 tweets de chaque thème (gender, religion et sexual orientation) nous avons observé quels sont les sujets du hate speech et comptabilisé leurs récurrences pour obtenir des pourcentages, de la même manière que pour les cibles.

GENDER

Dans ce thème, certains mots dans leur forme seule peuvent désigner quelque chose (**exemple** : pute = femme qui se prostitue) mais peuvent avoir une signification différente en contexte (**exemple** : t'es trop une pute de l'avoir tapé = ça ne se fait pas de l'avoir tapé, ce peut être un garçon ou une fille qui a fait l'action de taper). Dans ce cas, on analyse le mieux possible le contexte pour mieux classer le tweet mais si on ne trouve pas, si aucun indice ne peut nous aider à le catégoriser, on le place dans « autre ».

Les sujets de hate speech du thème "gender" sont :

- "tâche domestique": la cible est jugée sur sa capacité à faire une tâche domestique, que ce soit le ménage, la cuisine, la vaisselle...ou est désignée comme "assignée" à cette tâche.

Exemple: "La cuisine pour une femme EST UN DEVOIR NATUREL comme on parlerai de droit naturel." comme on peut clairement le voir dans ce tweet avec les majuscules, la cible, la femme, serait prédestinée à cuisiner. Elle n'aurait pas le choix de le faire.

- "sexualité": la cible est jugée sur sa sexualité, c'est-à-dire, avec qui, combien de fois… Cette catégorie comporte aussi les tweets ayant des insultes en rapport avec la sexualité : "pute", "salop(e)"...

Exemple : "Les femmes qui ont eu plus d'un partenaire sexuel comment pouvez-vous encore regarder les membres de vôtres famille…" ici, l'auteur juge les femmes qui ont eu plus d'un partenaire sexuel, comme si c'était une honte : "comment pouvez-vous encore regarder les membres de votre famille".

- "crédibilité": la cible est jugée sur sa crédibilité, la remise en cause de celle-ci par exemple.

Exemple : “Les femmes elles sont pas crédibles dans la démarche égalité homme/femme parce qu’elles se respectent déjà pas entre elle” ici, c’est explicite, l’auteur exprime hautement que “les femmes elles sont pas crédibles”.

- “physique” : le cible est jugée sur son physique, que ce soit sa tenue, sa beauté ou son corps…

Exemple : “Ministre et #EnMêmeTemps habillée comme une pute...” dans ce tweet la cible est clairement jugée sur sa tenue “habillée comme…” au point même d’être traitée de “pute”.

- "intelligence": la cible est jugée sur ses capacités cognitives, ou insultée de “conne”, “con”...

Exemple : “J’explique à une conne quelle est conne ça me dit que je fais de l’harcèlement” l’auteur juge ici clairement la cible sur ses capacités cognitives en l’insultant de “conne” deux fois de suite.

- “parole” : la cible est jugée sur ce qu’elle dit, son humour… Cette catégorie comprend également les insultes ordonnant à quelqu’un de se taire.

Exemple : “Ferme ta gueule connasse, toujours les mêmes qui l’ouvrent pour dire de la merde” on peut voir ici qu’à deux répétitions la femme est jugée sur sa parole, entre le moment où on lui ordonne de se taire et celui où on juge ce qu’elle a pu dire.

- “argent”: tout hate speech en lien avec l’argent, que ce soit la rémunération, l’attirance pour l’argent...

Exemple : “Pourquoi délocaliser alors que nous avons ici une main d'oeuvre pas chère... les femmes ?!” comme on peut le voir ici, le hate speech est en lien avec l’argent puisque la cible, les femmes en général, sont jugées comme “une main d'œuvre pas chère”.

- “objet sexuel”: la cible est jugée sur son physique de manière sexuelle, elle est perçue comme un objet sexuel par l’auteur.

Exemple : “T'as un cul à faire bander un aveugle.” La cible est ici clairement jugée physiquement sur son corps et de manière sexuelle avec le terme “bander”.

- “violence physique”: tweets explicitant des violences physiques subies, qu’elles soient sexuelles ou non. Il peut également y avoir des tweets en lien avec des agressions passées.

Exemple : “Campus d'université. Trop d'alcool. Aucun souvenir. Aucun consentement. Dur réveil auprès d'un inconnu. Lui, se souvenait.” La victime exprime clairement qu’elle a été violée : “aucun consentement”.

- “autre”: cette catégorie comprend les tweets ayant un sujet différent de ceux déjà cités ou des tweets où le sujet n’est pas identifiable.

Exemple : “Enceinte et sur le marché du travail? - c’est un handicap notoire” ici, le sujet de ce tweet est que la cible est enceinte et donc que ça pose problème, seulement ce sujet ne contenait pas assez de tweets pour en faire une catégorie, c’est pour cela qu’elle se trouve dans celle-ci.

- “2 sujets” : dans cette catégorie se trouvent les tweets pouvant entrer dans deux catégories existantes.
- **Exemple** : “Mon père, qui alors que je viens de me faire attoucher dans notre cage d'escalier m'a dit "t'as vu ta tenue?" J'avais 14 ans.”ici on peut voir que la cible a subi des violences physiques “viens de me faire attoucher” et qu’elle est jugée sur son physique “t’as vu ta tenue ?”.

Ce tweet peut donc aussi bien aller dans la catégorie “violence physique” que “physique”, or nous ne pouvons pas choisir de le mettre dans l’une de ces catégories plus que l’autre ni le mettre dans les deux puisque ce serait comme dupliquer le tweet, alors, pour que les chiffres de nos statistiques soient corrects, nous avons créé cette catégorie où les tweets sont analysés indépendamment des autres.

Document	Sujets	Récurrences	Total	Pourcentage
HS Gender	tâche domestique	6	200	3,0%
	sexualité	41		20,5%
	crédibilité	11		5,5%
	physique	26		13,0%
	intelligence	10		5,0%
	parole	6		3,0%
	argent	4		2,0%
	objet sexuel	53		26,5%
	violence physique	17		8,5%
	autre	19		9,5%
	2 sujets	7		3,5%

Comme nous pouvons le voir dans le tableau, le sujet le plus répandu de hate speech parmi ces 200 tweets est “objet sexuel” avec 26,5%, suivi de “sexualité” avec 20,5%. On peut donc dire que les cibles de hate speech sur le genre sont beaucoup sexualisées, c’est-à-dire jugées sur leur physique de manière sexuelle ou avec des intentions sexuelles, mais, elles sont aussi beaucoup jugées sur leur sexualité que ce soit à travers des insultes comme “pute” par exemple.

Pour ce qui est de la catégorie “autre”, nous avons observé chaque tweet de cette catégorie, auquel nous avons identifié le sujet si c’était possible, sinon, nous l’avons annoté comme “sujet non identifiable”, comme nous pouvons le voir dans un extrait du tableau global du thème Gender ci-dessous, lorsqu’un tweet est dans la catégorie “autre” nous avons précisé entre parenthèse sa “sous-catégorie”.

Cible	Cause	Ton humoristique	mots-clés
femme	tâche domestique	Non	cuisine + femme + devoir naturel + droit naturel
femme	tâche domestique	Non	femme + sait pas cuisiner + filles + pires qu'elle
femme	sexualité	Non	putain + vraiment des putes
femme	sexualité	Non	la plus grosse pute
femme	sexualité	Non	la pute de ta femme
femme	sexualité	Non	je m'en bats les couilles + il reste que + des putes
femme	autre (non identifiée)	Non	connasse par excellence
femme	crédibilité	Non	une autre connasse + elle pense savoir + mdr
femme	crédibilité	Non	dinde + #ladindesurvivante
femme	physique	Non	habillée comme une pute
femme	physique	Non	c'est + les femmes + auteures des harcèlements + habillements ultra-sexy
femme	autre (non identifiée)	Non	vraiment une connasse
femme	autre (non identifiée)	Non	la connasse
les deux	sexualité	Non	va t'occuper de ta putain de femme + arrête de jouer au foot
femme	intelligence	Non	elle reste plantée là comme une conne
femme	tâche domestique	Non	vous vous dites femmes + vous savez même pas faites des pâtes + bande de connasses
femme	crédibilité	Non	femmes + pas crédibles + se respectent pas entre elles
femme	physique	Non	espagnoles + boudins + #sexiste mais vrai
femme	parole	Oui	#blague + impossible + MOUAARF
femme	crédibilité	Non	exposition + femmes de foire + les potiches
femme	argent	Oui	femmes + mariraient + singes + #ironie

Ensuite, nous avons quantifié ces sujets pour avoir de nouveaux pourcentages. Seulement, il faut savoir qu'il y a plus de sujets que de tweets puisque certains peuvent en comporter plusieurs, c'est pourquoi nous avons choisi comme effectif total le nombre de sujets et pas le nombre de tweets de la catégorie, pour réaliser nos pourcentages.

Document	Sujet autre	Récurrences	Total	Pourcentage
HS Gender	non identifiée	9	23	39,0%
	hypocrisie	2		9,0%
	sport	2		9,0%
	action	1		4,0%
	conduite	1		4,0%
	enceinte	3		13,0%
	capacité	4		18,0%
	faible	1		4,0%

Nous pouvons donc voir qu'au sein de la catégorie "autre" de "gender", 39% des sujets de ces tweets "autres" sont non identifiables, c'est-à-dire que nous n'arrivons pas à déterminer pourquoi la cible subit du hate speech. Sinon, les deux autres catégories les plus répandues sont "capacité", c'est-à-dire, que la cible est jugée sur le fait de ne pas réussir ou ne pas pouvoir faire quelque chose, avec 17% et "enceinte" avec 13% où la cible est jugée sur le fait d'être enceinte, cela poserait problème à l'auteur ou à la personne qui harcèle.

Enfin, pour la catégorie "2 sujets", c'est-à-dire celle comprenant des tweets pouvant faire partie de deux catégories différentes, nous avons quantifié les catégories auxquelles ces tweets peuvent appartenir. Donc, comme pour la catégorie "autre", il y a plus de sujets que de tweets puisque chacun d'eux comporte deux sujets, c'est pourquoi, ici aussi, nous avons choisi comme effectif total le nombre de sujets et, pas le nombre de tweets de la catégorie pour réaliser nos pourcentages.

Document	Sujets	Récurrences	Total	Pourcentage
HS Gender	intelligence	2	14	14,3%
	physique	4		29,0%
	parole	1		7,0%
	tâche domestique	1		7,0%
	sexualité	2		14,3%
	objet sexuel	1		7,0%
	violence physique	2		14,3%
	autre	1		7,0%

Nous pouvons voir que le sujet le plus présent dans cette catégorie de tweets est “physique” avec 28% et donc que le physique est le sujet le plus répandu de hate speech dans le cas d’un tweet ayant plusieurs sujets. C’est-à-dire que les auteurs harcèlent souvent les victimes sur leur physique et autre chose en même temps. En effet, dans les 4 cas où un des sujets est le physique, l’autre est soit l’intelligence (1 cas), “violence physique” (2 cas) ou sexualité (1 cas). On peut remarquer que dans 3 cas sur 4 le physique est jugé de manière sexuelle et ça va plus loin, jusqu’à des agressions sexuelles dans 2 cas sur 4, ce qui n’est pas négligeable.

RELIGION

Pour ce thème, comportant des hashtags assez forts (= violents), nous avons choisi de placer tout tweet comportant un hashtag haineux envers une religion particulière, dans la catégorie “religion” puisque ce hashtag évoque une haine envers telle religion, sauf contre-indication, c’est-à-dire, sauf si le texte du tweet contredit le hashtag par exemple.

Les sujets de hate speech dans le thème “religion” sont :

- “physique”: dans le cas où les pratiquants sont jugés sur le physique ou sur quelque chose qu’ils doivent porter par exemple (voile, kippa, croix…).

Exemple : “#portduvoile arrêtons l’hypocrisie ! Faire dans le détail = + de problèmes ! IL FAUT INTERDIRE LE PORT DU VOILE EN FRANCE !!POINT .” On voit bien que le port du voile est le sujet de cette haine envers la cible, entre le hashtag, les phrases exclamatives montrant la conviction de l’auteur ainsi que la phrase toute en majuscules qui est on ne peut plus explicite pour classer ce tweet dans cette catégorie.

- “passé historique” : tout tweet faisant allusion ou citant un moment d’histoire douloureux pour une religion (**exemple** : la Shoah…).

Exemple : “#unbonjuif est un phœnix, bah oui ils renaissent de leur cendres” ici le tweet fait bien référence à la Shoah, puisque les cendres font référence aux personnes brûlées lors du génocide.

- “nourriture” : tout message lié au mode alimentaire d’une religion (**exemple**: nourriture halal, manger du porc…).

Exemple : “A Lyon, la mairie #EELV prive les écoliers de viande : c’est ce qui convient aux enfants qui ne mangent pas de porc. Imposer aux Français l'islamisation ! #Grandremplacement #migrants #DefendEurope #Remigration #StopIslam #FrontSocial #RIC” Dans ce tweet nous pouvons voir que la nourriture est le sujet du hate speech entre les hashtags explicites, la suite de mots “ne mangent pas de porc” et l’utilisation du mot “imposer”.

- “religion” : dans cette catégorie, nous trouvons les tweets qui visent une religion en général donc si un tweet comporte un texte neutre (= pas haineux) avec un hashtag haineux ou contre une religion (**exemple** : #christianophobie…), le tweet est catégorisé comme contre une religion donc, il entre dans cette catégorie.

Exemple : “Au moins c'est clair, le vrai problème c'est bien l'islam ! Contrairement à ce que nous disent les médias et les responsables politiques Français ! #Grandremplacement #migrants #DefendEurope #Remigration #StopIslam #FrontSocial #RIC” dans ce tweet il est dit “le vrai problème c’est bien l’islam” or on ne sait pas pourquoi l’islam est jugée comme un problème ici, donc nous pouvons dire que c’est la religion en soit qui est le sujet du hate speech ici.

- “lieux de culte” : catégorie qui comprend les tweets parlant des lieux de culte (mosquée, synagogue, église…) et de la profanation de ceux-ci.

Exemple : “#christianophobie ➖Les églises, usines, voitures qui brûlent ➖Les cimetières, statues saccagées ➖Ce sont des accidents ➖Qu’il disait ➖l'imbécile heureux @CCastaner #LREM ➖Les bouffées délirantes et les allumettes sont #mdr 🤣” dans ce tweet, l’auteur témoigne la haine envers les chrétiens à travers le hashtag “#christianophobie” et la citation de lieux religieux “saccagés”, “brûlés”. Il témoigne donc que ces violences physiques envers les églises sont dirigées vers l’Eglise, donc contre la religion catholique.

- “violence”: comprend des tweets violents ou liés à la violence.

Exemple : “Allah poubelle fout surtout la merde Haine, violence, attentats, soumission, ...Voilà le résultat du mal #StopIslamiste #StopIslam” dans ce tweet on peut voir que le sujet du hate speech est la violence (“haine”, “violence”, “attentats”, “soumission”).

- “migration” : messages de haine liés aux flux migratoires que ce soit avec un hashtag (**exemple**: #remigration) ou avec des mots en lien avec la migration (**exemple**: migrants…).

Exemple: “C'est faux, les Français sont ceux qui paient un maximum de taxes et d’impôts pour les autres et pour ne percevoir aucune aide pendant que l'immigration et les migrants sont bien au chaud...😠 #GrandRemplacement #Remigration #StopIslam #RIC #FrontSocial” ici, la sujet du hate speech est indiquée par l’association de la suite de mots “l'immigration et les migrants sont bien au chaud”, l’émoticône pas content montre l’état de l’auteur et le hashtag #remigration qui signifie “l’immigration retour”, le fait de “renvoyer les migrants chez eux”.

- “autre” : dans cette catégorie se trouvent les tweets ne rentrant pas dans une catégorie existante ou si le sujet est indéterminé.

Exemple : “Bientot nous accorderons + de crédit aux dealer et aux salafistes qu'aux gens polis & droits #bougnouls” ici nous voyons que l’islam est la cible avec le hashtag #bougnouls mais, le texte n’est pas “neutre” pour nous permettre de classer ce tweet dans la catégorie “religion”. Nous pouvons voir qu’il y a un sujet de haine seulement nous n’arrivons pas à la définir.

- “2 sujets” : cette catégorie comprend les tweets qui peuvent entrer dans deux catégories existantes.

Exemple : “C’est bien dommage !! Cette #France voilée on n’en veut plus, les mosquées 🕌 en première ligne non plus... Le manque de courage, l’aveuglement et le dogmatisme des politiques nous amènent droit à la guerre civile. #StopIslam #Stopauxvoiles #StopImmigration” nous pouvons voir que les sujets de hate speech de ce type sont le “physique” avec “voilée on n’en veut plus” et les “lieux de culte” avec “les mosquées [...] non plus”.

Document	Sujets	Récurrences	Total	Pourcentage
HS Religion	physique	18	200	9,0%
	passé historique	12		6,0%
	nourriture	5		2,5%
	religion	81		40,5%
	lieux de culte	21		10,5%
	violence	35		17,5%
	migration	21		10,5%
	autre	3		1,5%
	2 sujets	4		2,0%

Comme nous pouvons le voir dans le tableau ci-dessus, le sujet le plus répandu de hate speech dans le thème “religion” est “religion” avec 40,5%, ce qui signifie que ce n’est pas une chose imposée par la religion (**exemple** : porter le voile, ne pas manger de porc…) qui gêne le plus les auteurs de ce type de tweets mais la religion en elle-même.

Document	Sujet autre	Récurrences	Total	Pourcentage
HS Religion	prénom	1	3	33,3%
	respect	1		33,3%
	non identifiée	1		33,3%

Nous pouvons voir que chacun des sujets est autant représenté qu'un autre dans cette catégorie. En effet, chacun des trois tweets présents dans la catégorie “autre” présente un sujet différent. Il n’y en a donc pas un qui apparaît plus qu'un autre.

Document	Sujets	Récurrences	Total	Pourcentage
HS Religion	physique	1	8	12,5%
	lieux de culte	4		50,0%
	violence	2		25,0%
	religion	1		12,5%

Pour la catégorie “2 sujets” de ce thème, nous pouvons voir que “lieux de culte” est le sujet le plus répandu dans cette catégorie. En effet, ce sujet apparaît dans chaque tweet et est associé au sujet physique (1 cas), à la violence d’une religion (2 cas) ou à la religion en général (1 cas). On peut donc voir que dans 2 cas sur 4, soit 1 tweet sur 2 de cette catégorie comporte du hate speech sur les lieux de culte et de la violence. Donc, dans cette catégorie, la violence physique ou verbale semble liée aux lieux de culte.

SEXUAL ORIENTATION

Les sujets de hate speech du thème “sexual orientation” sont les suivants :

- “orientation sexuelle” : les tweets jugeant l’orientation sexuelle de la cible.

Exemple : “#simonfilsestgay mdrrr c plus mon fils” on peut voir que la personne pourrait rejeter son fils à cause de son attirance pour les hommes.

- “nature” : jugement sur l’orientation sexuelle de la cible par rapport à la nature, la reproduction, la religion...

Exemple : “#LesGaysdoiventdisparaitrecar C'EST CONTRENATURHAN” on peut voir ici que la cible subit du hate speech parce que l’auteur juge que son orientation sexuelle “contre nature”.

- “pratique sexuelle” : jugement lié à la pratique sexuelle de la cible.

Exemple : “#LesGaysDoiventDisparaîtreCar la sodomie ca fait mal ^^” on voit bien ici un jugement sur la pratique sexuelle de la cible avec le terme “sodomie”.

- “infériorité” : comme nous l’avons dit précédemment, les hashtags ou mots “tantouze”, “tapette” peuvent désigner un homme homosexuel mais également quelqu’un qui a peur, ou quelqu’un d’inférieur. Dans le cas d’un tweet où un de ces termes est utilisé sans indice qui indique la désinence d’un homosexuel, le tweet sera catégorisé ici, sauf contre-indication.

Exemple : “#BrulonsLesGaysSurDu Bois spécial sous race” ici l’utilisation de “sous race” montre bien que la cible est jugée comme inférieure.

- “autre” : dans cette catégorie se trouvent les tweets ne rentrant pas dans une catégorie existante ou si le sujet est indéterminée.

Exemple : “#BrulonsLesGaysSurDu kaaris - zoo "les gays viennent de sortir du zoo”” ici nous pouvons voir qu’il y a un sujet de hate speech mais nous ne parvenons pas à déterminer lequel. Les gays sont-ils ici vus comme des animaux (“zoo”) ?

- “2 sujets” : cette catégorie comprend les tweets qui peuvent entrer dans deux catégories existantes.

Exemple : “C’est une femme @neymarjr il s’est fait greffer un gland et basta !! #tafiole” on peut voir que ce tweet peut aussi bien aller dans la catégorie “autre” puisque nous pouvons voir que la femme, donc le sexe féminin est jugé, ainsi que dans la catégorie “infériorité” étant donné la présence de l’hashtag #tafiole.

Document	Sujets	Récurrences	Total	Pourcentage
HS Sexual Orientation	orientation sexuelle	137	200	68,5%
	nature	14		7,0%
	pratique sexuelle	14		7,0%
	infériorité	27		13,5%
	autre	7		3,5%
	2 sujets	1		0,5%

Ici, nous pouvons voir que le sujet de hate speech le plus représenté dans ce thème est “sexualité” avec 68,5%. Ce qui signifie que ce qui “gêne” le plus les auteurs de ces tweets est le fait que les cibles soient attirées par quelqu’un du même sexe de manière générale et pas quelque chose en particulier comme la pratique sexuelle par exemple.

L’autre sujet le plus représenté est “infériorité” où on peut voir que 13,5% de ces tweets jugent une personne homosexuelle sur ce sujet. Mais, ces tweets peuvent aussi bien juger “sérieusement” la cible comme inférieure ou de manière ironique.

Document	Sujet autre	Récurrences	Total	Pourcentage
HS Sexual Orientation	sexe	1	7	14,0%
	non identifiée	4		58,0%
	intelligence	1		14,0%
	exhibitionnisme	1		14,0%

Dans la catégorie “autre” de ce thème, nous pouvons voir que 57% des sujets de ces tweets ne sont pas identifiés et donc qu’il y a bien du hate speech envers quelqu’un mais nous ne savons pas pourquoi.

Document	Sujets	Récurrences	Total	Pourcentage
HS Sexual Orientation	autre	1	2	50,0%
	infériorité	1		50,0%

Dans ce thème, il n’y a qu’un seul tweet pouvant être dans deux catégories différentes. Il peut aussi bien être dans “autre” que dans “infériorité” donc l’infériorité de la cible serait associée à autre chose que nous ne parvenons pas à identifier.

3. Analyse des corrélations entre les catégories

Dans cette partie de l'analyse, nous allons comparer les sujets et les cibles de hate speech pour voir si nous observons des récurrences. Par exemple, si nous observons qu'une cible est plus victime de hate speech sur un sujet qu'une autre ou s'il y a des sujets récurrents entre les thèmes.

Nous avons choisi d'observer les cibles parmi les sujets et pas l'inverse puisque cela nous permet d'avoir plus de précisions et c'est plus intéressant. Pour se faire, nous avons fait un tableau entre celui des cibles et celui des sujets. Nous avons regardé parmis les différents sujets de hate speech, quelles sont les cibles les plus récurrentes, ce qui nous a donné ceci :

Document	Sujets	Récurrences	Cibles	Récurrences	Pourcentage
HS Gender	tâche domestique	6	femmes	6	100,0%
	sexualité	41	femmes	37	90,0%
			hommes	1	3,0%
			les deux	3	7,0%
	crédibilité	11	femmes	10	91,0%
			cible non définie	1	9,0%
	physique	26	femmes	24	92,0%
			hommes	1	4,0%
			cible non définie	1	4,0%
	intelligence	10	femmes	9	90,0%
			hommes	1	10,0%
	parole	6	femmes	4	66,0%
			hommes	1	17,0%
			cible non définie	1	17,0%
	argent	4	femmes	4	100,0%
	objet sexuel	53	femmes	49	92,0%
			cible non définie	4	8,0%
	violence physique	17	femmes	10	59,0%
			cible non définie	7	41,0%
	autre	19	femmes	16	84,0%
			les deux	1	5,0%
			cible non définie	2	11,0%
	2 sujets	7	femmes	5	72,0%
			les deux	1	14,0%
			cible non définie	1	14,0%

Document	Sujets	Récurrences	Cibles	Récurrences	Pourcentage
HS Religion	physique	18	Judaïsme	1	6,0%
			Islam	17	94,0%
	passé historique	12	Judaïsme	12	100,0%
	nourriture	5	Judaïsme	1	20,0%
			Islam	4	80,0%
	religion	81	Judaïsme	11	14,0%
			Islam	61	75,0%
			Christianisme	7	9,0%
			cible non définie	2	2,0%
	lieux de culte	21	Islam	2	10,0%
			Christianisme	19	90,0%
	violence	35	Judaïsme	3	9,0%
			Islam	22	63,0%
			Christianisme	6	17,0%
			cible non définie	4	11,0%
	migration	21	Islam	18	86,0%
			Christianisme	1	5,0%
			cible non définie	2	9,0%
	autre	3	Islam	3	100,0%
	2 sujets	4	Islam	1	25,0%
			Christianisme	2	50,0%
			cible non définie	1	25,0%

Document	Sujets	Récurrences	Cibles	Récurrences	Pourcentage
HS Sexual Orientation	orientation sexuelle	137	hommes	23	17,0%
			femmes	12	9,0%
			les deux	3	2,0%
			cible non définie	99	72,0%
	nature	14	hommes	4	29,0%
			cible non définie	10	71,0%
	pratique sexuelle	14	hommes	10	71,5%
			femmes	1	7,0%
			cible non définie	3	21,5%
	infériorité	27	hommes	24	88,0%
			femmes	1	4,0%
			les deux	1	4,0%
			cible non définie	1	4,0%
	autre	7	hommes	1	14,0%
			cible non définie	6	86,0%
	2 sujets	1	hommes	1	100,0%

Ces trois tableaux nous permettent de voir quelle cible est la plus visée selon le sujet. Dans le thème “Gender”, nous pouvons voir que les femmes sont toujours les cibles dans le cas de tweets portants sur les tâches ménagères ou sur l’argent et qu’elles sont visées en très grande majorité (> 75%) dans le cas de tweets sur la sexualité, la crédibilité, le physique, l’intelligence ou des tweets ayant pour sujet “objet sexuel” et “autre”.

Pour ce qui est des autres catégories, on ne peut pas dire que ce sont les hommes qui sont visés puisqu'ils ne sont pas explicitement cités. En effet, l'autre cible qui est aussi assez souvent citée est "cible non définie", c'est-à-dire que nous ne pouvons pas savoir si ce sont les hommes, les femmes ou les deux sexes qui sont visés dans ces cas-là.

Dans le thème "Religion", il apparaît que les juifs sont toujours les cibles dans le cas de tweets en lien avec le passé historique, l'Islam est toujours la cible pour les tweets présents dans la catégorie "autre" et est visée en très grande majorité (> 75%) dans le cas de tweets ayant pour sujet le physique, la nourriture, la religion et la migration. Le Christianisme, lui, est en très grande majorité visé lors de tweets portant sur les lieux de culte.

Nous pouvons donc assez nettement voir que l'Islam est la religion qui subit le plus le hate speech, dans une majorité de sujets et qu'elle est entre autres jugée sur des choses qui sont "imposées" par la religion même, comme le fait que les femmes doivent porter un voile (physique), manger du porc n'est pas autorisé (nourriture) etc.

Enfin, les hommes sont la cible en grande majorité dans le cas de tweets sur l'infériorité et la pratique sexuelle. Donc les homosexuels sont plus jugés comme "inférieurs" que les femmes que ce soit dans le sens "sérieux" où ils sont considérés comme "sous-hommes" ou dans le sens plus ironique (on ne sait pas vraiment si c'est sérieux ou pas) où ils sont jugés parce qu'ils ont peur (**exemple** : #tapette).

Dans la catégorie "autre", la cible la plus visée est "cible non définie" nous ne pouvons pas vraiment dire qu'un sexe est plus visé qu'un autre. Ce qui est aussi le cas dans la plupart de ce thème, puisqu'une très grande partie des cibles est "non définie" et donc, nous ne pouvons pas vraiment savoir ici, qui est le plus visé, les hommes ? Les femmes ? Les deux ?

Nous avons choisi d'effectuer une autre analyse entre les cibles et les sujets qui consiste à observer quels sont les sujets de hate speech en commun aux trois thèmes. Pour cela, nous avons choisi d'utiliser les tableaux créés dans la première partie et nous avons aussi pris en compte les sous-catégories dans les sujets "autre" de chaque thème.

Nous pouvons voir que la catégorie “autre” et “2 sujets” sont évidemment communes aux trois thèmes puisque nous avons procédé de la même manière pour trier nos données, dans chacun des thèmes.

Certains sujets en rapport avec tout ce qui est verbal (“intelligence”, “hypocrisie”, “parole”, “respect”) sont aussi communs aux trois thèmes. Nous pouvons donc dire que peu importe le thème, les cibles sont également jugées sur ce qu’elles disent ou pensent.

Nous observons que le sujet “violence physique” de “Gender” et “violence” de “Religion” sont semblables, en effet, comme leur nom l’indique, ils font tous les deux références à des faits violents (sexuels, physiques ou verbaux). Ceci nous laisse penser que ce sont deux thèmes en rapport avec la violence mais d’une manière différente : le thème Gender va plutôt rapporter de la violence sexuelle, tandis que “Religion” traîte de violence physique ou verbale.

Dans ces deux mêmes thèmes nous retrouvons aussi le sujet “physique”, ce qui signifie que les cibles de ces deux thèmes subissent aussi du hate speech envers leur physique, notamment sur leur manière de s’habiller (tenue courte, sexy ou port du voile, kippa…).

Entre les thèmes “Gender” et “Sexual Orientation”, nous retrouvons plusieurs sujets comme “sexualité”, “orientation sexuelle” et “pratique sexuelle” qui peuvent s’apparenter au sexe mais qui permettent de couvrir les différents aspects de celui-ci. Cela nous permet tout de même de constater que le sexe est un thème très jugé dans “Gender” et “Sexual Orientation”, mais pas du tout dans “Religion”, sûrement parce que les deux premiers thèmes sont liés ou assez proche de la sexualité tandis que ce n’est pas du tout le cas de la religion, au contraire, le sexe y est même un peu tabou.

4. Analyse du ton humoristique

L’humour est défini comme étant une “forme d’esprit qui consiste à dégager les aspects plaisants et insolites de la réalité, avec un certain détachement”, selon le dictionnaire Le Robert. Il en existe plusieurs formes : l’ironie, la blague… Utiliser ce procédé permet de semer le doute dans l’esprit de l’interlocuteur, puisqu’il ne saura pas vraiment si le locuteur pense vraiment ce qu’il vient de dire ou non.

Pour pouvoir répertorier les tweets ayant un ton humoristique, il nous a suffi de repérer dans chacun d'eux des indices comme des hashtags, des mots, des émoticônes : "ahah", "#ironie", "#blague", "^^", mdr (= mort de rire) etc. Ces mots permettent également de nous indiquer quel type d'humour il s'agit : la présence de l'hashtag #ironie nous montre que ce tweet est ironique, un "ahah" nous montre le ton blagueur de l'auteur...

Exemple : "aujourd'hui j'ai vu à quel point l'amitié entre fille pouvait être sincère et vraie, ahah #ironie" dans ce tweet, nous pouvons bien voir qu'il s'agit d'humour et plus précisément d'ironie, grâce à l'hashtag mais aussi parce que l'ironie est définie comme étant "une manière de s'exprimer (de quelqu'un ou de quelque chose) en disant le contraire de ce que l'on pense", et c'est exactement ce que nous avons ici, nous ne pouvons pas déterminer si l'auteur de ce tweet est sérieux.

La présence d'indices comme nous venons de le voir peut nous permettre de classer les tweets comme humoristique, sauf s'il y a une contre-indication. Par exemple dans le tweet "encore une autre connasse qui veut être tatoueur parce qu'elle pense savoir dessiner ms mdr" ici, le "mdr" indique plutôt "elle pense savoir dessiner mais c'est faux", il n'indique pas d'humour. Un tweet n'est également pas considéré comme humoristique si l'indice se trouve en commentaire d'un discours rapporté.

Exemple : ""@nabilaboss: #BrulonsLesGaysSurDu elton john ! On est gentil on vs brule sur un chanteur gay" PTDRRRRRR" le ton du tweet rapporté n'est pas humoristique, le "ptdrrr" (= pété de rire) est ici utilisé comme un commentaire à ce message.

Ensuite, il nous a suffi de comptabiliser les tweets de chaque thème étant humoristiques et de réaliser un tableau pour pouvoir réaliser des pourcentages.

Document	Ton humoristique	Récurrences	Total	Pourcentage
HS Gender	Oui	18	200	9,0%
	Non	182		91,0%
HS Religion	Oui	1	200	0,5%
	Non	199		99,5%
HS Sexual Orientation	Oui	31	200	15,5%
	Non	169		84,5%

Ce tableau nous permet de remarquer que le thème qui comporte le plus de tweets avec un ton humoristique est “Sexual Orientation” avec 15,5% contre 9% pour “Gender” et 0,5% pour “Religion”.

On peut donc dire que les auteurs des tweets “Sexual Orientation” ont plus tendance à utiliser l’humour pour faire du hate speech que ceux des autres thèmes. Et que, les auteurs des tweets sur la religion vont très peu en faire, voire pas du tout (seulement 1 tweet sur 200), peut-être parce que ce thème est plus sensible, plus “sérieux” et donc, ils ne se permettent pas de faire de blague pour dire ce qu’ils pensent, ils vont droit au but.

Dans cette deuxième partie de l’analyse du ton humoristique, nous allons voir quelles sont les différentes formes d’humour utilisées et leurs proportions dans les trois corpus. Pour se faire, nous avons répertorié les tweets annotés comme “humoristiques” dans un document, puis nous avons précisé à quelle forme d’humour chacun appartient (ironie, blague…) et compté leurs récurrences. Pour ce qui est d'annoter quel type d’humour le tweet comporte, nous avons choisi de considérer qu’un tweet avec l’hashtag #ironie appartient à la forme ironique, d’autant plus si nous constatons que l’auteur semble dire l’inverse de ce qu’il pense. Les tweets avec l’hashtag #blague sont considérés comme appartenant à la classe “blague”, ce peut être aussi des tweets ne comportant pas ce hashtag mais où il est explicite que l’auteur fait une blague (**exemple** : “#BrulonsLesGaysSurDu putain mais srx vous allez trop loin là ... Mdrrr nan jrigole niquez biens vos pères”). Les autres tweets sont considérés comme “humour” puisque nous ne pouvons pas être sûrs que c’est une blague ou de l’ironie, il peut aussi s’agir de tweets tout à fait sérieux avec un “indice d’humour” pour que le message “passe mieux”, pour qu’il soit moins mal pris.

Dans tous les cas, l’humour est utilisé par l’auteur pour atténuer le message qu’il veut faire passer et donc, il se "cache" derrière son discours pour ne pas se retrouver en tant que “harceleur direct”. De plus, l’interlocuteur de ce genre de tweets sera moins touché par ces paroles s’il n’a pas connaissance de l’exactitude de ce qui est dit.

Document	Ton humoristique	Récurrences	Total	Pourcentage
HS Gender	ironie	16	18	89,0%
	blague	1		5,5%
	humour	1		5,5%
HS Religion	ironie	1	1	100,0%
	blague	0		0,0%
	humour	0		0,0%
HS Sexual Orientation	ironie	0	31	0,0%
	blague	2		6,0%
	humour	29		94,0%

Nous obtenons donc ce tableau, qui nous permet de voir que les tweets sont principalement humoristiques (30 tweets sur 50) ou ironiques (17 tweets sur 50). Le thème “Gender” comporte 89% de tweets ironiques parmi ses tweets avec un ton humoristique, “Religion” 100% d’ironie (mais ne comporte qu’un seul tweet au total) et “Sexual Orientation” comprend 94% de tweets humoristiques.

Selon un article de Brigitte Bouquet et Jacques Riffault, l’ironie est une forme plus “méchante”, “dure” que l’humour. Nous pouvons donc penser que les auteurs des tweets ironiques de “Gender” se permettent d’utiliser l’ironie dans ce thème puisqu’il est plus “léger” que “Sexual Orientation” et que leurs propos dans ce thème, ne feraient pas autant polémique que si l’ironie était utilisée dans “Sexual Orientation”.

II. Partie 2 : Expérimentation et analyse des données recueillies

Dans le cadre d’une démarche de prévention et d’analyse du cyber-harcèlement, nous avons réalisé une expérimentation au sein d’un lycée pour pouvoir recueillir des données types mais aussi pour ensuite sensibiliser les élèves sur les conséquences du cyber-harcèlement.

A. Mise en place de l’expérimentation

Pour réaliser cette expérimentation, des scénarios ont été créés par un sociologue membre du projet OTESIA pour aider les élèves à contextualiser l’histoire dans laquelle ils allaient jouer et identifier le rôle de leur personnage. Ces scénarios portent sur différents thèmes comme “religion”, “ethnicité”, “homophobie” et “surpoids”. Voici ci-dessous le scénario du thème “Surpoids 1”.

Scenario SURPOIDS 1

Brice est un garçon timide avec quelques kilos en trop. Pendant un voyage scolaire à Prague, il partage la chambre avec des camarades de classe, dont David, le garçon le plus populaire de l'école, Amine, le meilleur ami de David, et Yanis. David réussit à prendre une photo de Brice pendant qu'il prend sa douche et la partage à toute la classe. Après un moment, David, avec l'aide d'Amine et de Jim, commence à se moquer de Brice. Yanis, qui a assisté à l'histoire, défend Brice sur le tchat avec le soutien de Moussa. Jules, ami à la fois de Brice et de David, intervient pour que le harcèlement cesse entre Brice et David.

Ensuite nous avons assigné les personnages au rôle qu'ils jouent dans l'histoire, pour chaque scénario il y avait au moins un harceleur avec un soutien de l'harceleur, une victime et au moins un soutien ainsi qu'un conciliateur qui permettait de temporiser la conversation pour éviter que ça aille trop loin. Nous pouvons voir ci-dessous la répartition des personnages et leur rôle pour "Surpoids 1".

- Scénario Surpoids 1 **(6 étudiants)**
  - 0 fille
  - 6 garcons
  - 1 victime
-

Harceleur	David (g)
Défenseur/Soutien de la victime	Yanis (g) Moussa (g)
Soutien du harceleur	Amine (g) Jim (g)
Victime	Brice (g)
Conciliateur	Jules (g)

Ensuite, nous avons donné à chaque élève, placé sur un ordinateur, un rôle. Aucun élève n'avait le rôle de victime, ce sont des personnes qui travaillent avec l'équipe OTESIA qui ont joué les personnages victimes, par sécurité pour les élèves. Une fois que chaque personne avait un rôle, les personnes du même scénario devaient se connecter au chat correspondant et "discuter" sur une plateforme en ligne, grâce à laquelle nous avons ensuite récupéré les messages échangés pour les analyser. Ce qui nous a donné ce que l'on voit ci-dessous, un extrait de la conversation sur le thème "Surpoids 1".

```
[2:28:36 PM] <David-Harceleur> alors la con de ta mère
[2:28:45 PM] <Jim-Soutien-Harceleur> OUE GROSPORC
[2:29:05 PM] <Jules-Conciliateur> vous etes mechant
[2:29:05 PM] <Jim-Soutien-Harceleur> il est ou le mamouth
[2:29:10 PM] <Amine-Soutien-Harceleur> IL EST OU LE GROS LARD
[2:29:25 PM] <David-Harceleur> espece de boule je te pousse tu roule
[2:29:38 PM] <Jules-Conciliateur> arretez c'est pas cool
[2:29:43 PM] <Jim-Soutien-Harceleur> et oh baleine bleu
[2:29:45 PM] <Amine-Soutien-Harceleur> ALLEZ MANIFESTE TOI GROS PORCS
[2:31:11 PM] <Jim-Soutien-Harceleur> y prl pas se chien
[2:31:16 PM] <Amine-Soutien-Harceleur> brice le cachalot
[2:31:22 PM] <David-Harceleur> vas courir gros phoque retourne dans l'eau
[2:31:33 PM] <Jules-Conciliateur> sa pourrais etre vous avec les kilos en trop
[2:31:41 PM] <Amine-Soutien-Harceleur> ta mere la reine des putes
[2:31:46 PM] <Jules-Conciliateur> david c pas bien wesh
[2:32:16 PM] <Jim-Soutien-Harceleur> merci au  client fidele
[2:32:32 PM] <David-Harceleur> aller les gros
[2:32:36 PM] <Jim-Soutien-Harceleur> alllez les gros
[2:32:54 PM] <Amine-Soutien-Harceleur> ton  corps il contient 90% de tartiflette sale chien
```

Une fois toutes les conversations recueillies, nous les avons nettoyées, c'est-à-dire que nous avons enlevé les messages inutiles tels que les "bonjour" ou "au revoir", par exemple pour ensuite passer à l'analyse. Au total, nous avons huit conversations dont deux sur le thème de l'ethnicité, deux sur l'homophobie, une sur le thème de la religion puis trois sur le thème du surpoids. Chaque conversation a une longueur différente en fonction de la réactivité des participants : "Ethnicité 1" comporte 203 messages, "Ethnicité 2" 43, "Homophobie 1" comporte 94 messages contre 106 pour "Homophobie 2", "Religion 2" comporte 173 échanges, "Surpoids 1" en a 132, "Surpoids 2" 179 et "Surpoids 2a" en comporte 82.

Puisque nous avons annoté chaque intervention de chaque conversation comme ayant ou non du hate speech, nous pouvons calculer la proportion de messages de haine dans ces conversations et remarquer dans le tableau ci-dessous que la proportion de message de haine varie d'une conversation à une autre. En effet, la proportion n'est pas la même entre les deux conversations sur le thème de l'ethnicité par exemple. On peut voir que "Ethnicité 1" comporte 46% de messages de haine, donc même pas la moitié alors que "Ethnicité 2" en comporte 77%. Pour le thème de l'homophobie, les deux conversations ont environ 60% de messages de haine. "Religion 2" comporte 45% de hate speech, et pour les conversations sur le surpoids, c'est autour de 50% pour "Surpoids 1" et "Surpoids 2" mais 63% pour "Surpoids 2a".

Thème	Analyse	Valeur	Récurrence	Pourcentage
Ethnicité 1	Hate Speech	oui	94	46%
		non	109	54%
Ethnicité 2	Hate Speech	oui	33	77%
		non	10	23%
Homophobie 1	Hate Speech	oui	62	66%
		non	32	34%
Homophobie 2	Hate Speech	oui	65	61%
		non	41	39%
Religion 2	Hate Speech	oui	78	45%
		non	95	55%
Surpoids 1	Hate Speech	oui	69	52%
		non	63	48%
Surpoids 2	Hate Speech	oui	99	55%
		non	80	45%
Surpoids 2a	Hate Speech	oui	52	63%
		non	30	37%

B. Analyse des données recueillies

Pour pouvoir analyser les conversations, il nous a fallu annoter chaque intervention, c'est-à-dire chaque ligne pour chaque conversation en reprenant le même schéma d'annotation des messages relevés sur Twitter (partie 1 du Mémoire). Est-ce que ce message est un message de haine ? Si oui, quelle est la cible ? Le sujet ? Y a-t-il un ton humoristique ? Le message de haine est-il formulé de manière implicite ? Nous pouvons voir un extrait de ces annotations pour "Surpoids 1" ci-dessous.

LIGNE	HS	CIBLE	FONCTION	SUJET	'ON HUMOUI	INDIRECTE	DETAIL
1	oui	brice	victime	physique	non		
2	oui	brice	victime	physique	non	oui	gros porc = gros
3	non						
4	oui	brice	victime	physique	non	oui	mamouth=gros
5	oui	brice	victime	physique	non	oui	gros lard = gros
6	oui	brice	victime	physique	non	oui	boule, tu roules = gros
7	non						
8	oui	brice	victime	physique	non	oui	baleine bleue = gros
9	oui	brice	victime	physique	non	oui	gros porc = gros
10	oui	brice	victime	autre	non	oui	chien = batard
11	oui	brice	victime	physique	non	oui	cachalot = gros
12	oui	brice	victime	physique	non	oui	gros phoque = gros
13	non						
14	oui	jules	conciliateur	autre	non		
15	non						
16	non						
17	non						

Grâce à ces annotations, nous allons pouvoir analyser ces conversations et voir quelle cible est la plus visée, les causes de messages de haine les plus récurrentes dans les thèmes, la proportion de messages avec un ton humoristique et si les auteurs insultent de manière directe ou implicite.

1. Analyse des données “cibles”

Dans cette analyse, nous avons procédé de la même manière que pour les cibles de tweets relevés dans la partie 1 : nous avons identifié les cibles de chaque message, plus précisément leur fonction dans le scénario (victime, harceleur, soutien de la victime (svictime), soutien de l’harceleur (sharceleur), conciliateur…) puisque c’est ce qui nous intéresse le plus ici, compté les récurrences et calculé les pourcentages, ce qui nous a donné ce tableau pour tous les thèmes.

Thème	Analyse	Valeur	Récurrence	Pourcentage
Ethnicité 1	Fonction sujet	victime	47	50%
		svictime	11	12%
		harceleur	12	13%
		sharceleur	8	8%
		non identifiée	10	11%
		conciliateur	6	6%
Ethnicité 2	Fonction sujet	victime	12	37%
		svictime	3	9%
		harceleur	7	21%
		sharceleur	2	6%
		non identifiée	4	12%
		extérieur	5	15%
Homophobie 1	Fonction sujet	victime	43	69%
		sharceleur	4	6%
		harceleur	1	2%
		conciliateur	5	8%
		non identifiée	8	13%
		victime + conciliateur	1	2%
Homophobie 2	Fonction sujet	victime	41	63%
		harceleur	6	9%
		svictime	2	3%
		sharceleur	2	3%
		conciliateur	1	2%
		non identifiée	12	18%
		victime + conciliateur	1	2%
Religion 2	Fonction sujet	harceleur	32	41%
		victime	21	27%
		sharceleur	6	8%
		svictime	4	5%
		non identifiée	14	18%
		conciliateur	1	1%
Surpoids 1	Fonction sujet	harceleur	3	4%
		victime	35	51%
		conciliateur	14	21%
		svictime	3	4%
		non identifiée	9	13%
		sharceleur	5	7%
Surpoids 2	Fonction sujet	victime	40	41%
		harceleur	19	19%
		svictime	11	11%
		non identifiée	26	26%
		sharceleur	3	3%
Surpoids 2a	Fonction sujet	victime	24	46%
		harceleur	20	38%
		svictime	1	2%
		non identifiée	6	12%
		sharceleur	1	2%

Pour le thème “Ethnicité 1”, nous pouvons voir que la moitié des messages de haine visent la victime elle-même et 12% les soutiens de la victime, donc le groupe des victimes est visé à 62% ce qui est très conséquent. De plus 11% sont des cibles non identifiées, ce peut être les victimes ou les harceleurs nous ne savons pas, 6% des messages visent le conciliateur, alors, les 21% restant visent les harceleurs. Donc, nous pouvons voir que les victimes se sont quand même défendues mais elles se sont beaucoup plus fait harceler. Dans “Ethnicité 2”, les victimes se sont moins fait harceler (46%) en tout et ont plus harcelé (27%) que dans la conversation précédente sur le même thème. Des personnes extérieures à la conversation ont également été visées ici (15%). On remarque également que l’harceleur est la deuxième personne la plus harcelée dans cette conversation aussi. Donc la victime a essayé de se défendre en harcelant à son tour l’harceleur.

“Homophobie 1” comprend 69% de messages qui visent la victime contre 8% contre les harceleurs, on peut donc dire que la victime ne s’est quasiment pas défendue et s’est fait harceler tout le long des échanges. Dans l’autre conversation sur ce thème (“Homophobie 2”), c’est le même schéma : la victime subit le harcèlement à 63%, ses soutiens aussi (3%) mais le groupe victime semble s’être plus défendu que dans la conversation précédente puisque le groupe de harceleurs est visé à 12%. Le conciliateur est aussi moins visé dans cette conversation que dans “Homophobie 1”, donc les membres se sont moins acharnés sur lui qui essayait de temporiser la conversation.

Dans la seule conversation sur le thème religion (“Religion 2”), nous observons que le groupe de victimes est visé à 32% contre 49% pour le groupe de harceleurs, ce qui est étonnant. Cela signifie que le scénario s’est en quelque sorte inversé, la victime semble être devenue harceleur et vice versa. Le conciliateur est aussi beaucoup moins visé dans cette conversation que dans toutes les précédentes, alors soit ils l’ont ignoré, soit ils l’ont très peu harcelé.

Le thème surpoids compte lui 3 conversations, la première (“Surpoids 1”) comprend 51% de messages qui visent uniquement la victime contre seulement 4% qui visent uniquement l’harceleur. Les soutiens de la victime ont été visés à 4% aussi et ceux de l’harceleur à 7%. Le conciliateur a été visé à 21%.

On peut donc dire que la victime est la cible dans plus d'un message de haine sur deux et que les soutiens de l'harceleur ont plus été harcelés que l'harceleur lui-même, soit parce que l'harceleur n'a pas trop harcelé et donc on ne l'a pas trop visé en retour, soit parce que les victimes s'en sont plutôt pris aux soutiens qu'au harceleur lui-même. De plus, le conciliateur a, ici, beaucoup subi de hate speech aussi avec 21% des messages de cette conversation, il est donc la deuxième personne la plus visée dans ces messages. "Surpoids 2" comprend 52% des messages qui visent les victimes contre 22% qui visent les harceleurs et 26% pour lesquels nous ne savons pas qui ils visent.

Les victimes semblent ici bien plus s'être défendues que précédemment, mais il y a aussi deux fois plus de messages pour lesquels nous ne savons pas quelles sont les cibles. Donc ce n'est pas vraiment figé comme résultat puisque selon les cibles non identifiées, cela pourrait faire pencher la balance. Enfin, dans la dernière conversation ("Surpoids 2a"), nous pouvons voir que la victime et l'harceleur sont presque autant visés l'un que l'autre à quatre messages près, ce qui montre que la victime s'est soit très bien défendue, soit elle a bien été défendue par ses soutiens.

D'une manière générale nous pouvons remarquer que les victimes sont bien les cibles des messages de haine, mais que dans les conversations portant sur le thème de l'homophobie, elles le sont plus que les autres et que dans la conversation sur le thème religion, il y a eu un changement de situation, puisque nous pouvons voir que l'harceleur est plus visé que la victime. Pour ce qui est des conciliateurs, ils sont peu visés, sauf dans la première conversation sur le surpoids où ils le sont à 21% contre 5% en général.

2. Analyse des données "sujets"

Pour pouvoir analyser les sujets, nous avons repéré pour chaque message la cause du hate speech que nous avons nommée, puis nous avons regroupé les sujets semblables, compté les récurrences et effectué les pourcentages.

Thème	Analyse	Valeur	Récurrence	Pourcentage
Ethnicité 1	Sujets	argent	2	2%
		physique	11	12%
		2 sujets	3	3%
		violence	9	10%
		autre	14	15%
		parole	17	18%
		origine	16	17%
		sexualité	8	9%
		éducation	2	2%
		infériorité	6	6%
		intelligence	5	5%
		jalousie	1	1%
Ethnicité 2	Sujets	dragué	4	12%
		sexualité	7	21%
		2 sujets	2	6%
		autre	8	25%
		violence	2	6%
		physique	6	18%
		origine	3	9%
		infériorité	1	3%
Homophobie 1	Sujets	violence	17	27%
		autre	10	16%
		orientation sexuelle	2	3%
		sexualité	6	10%
		parole	10	16%
		maladie	4	6%
		pas normal	1	2%
		infériorité	3	5%
		intelligence	4	6%
		physique	1	2%
		origine	1	2%
		sale	3	5%
Homophobie 2	Sujets	sale	16	24%
		parole	9	14%
		sexualité	11	17%
		autre	6	9%
		orientation sexuelle	16	24%
		nature	1	2%
		maladie	4	6%
		violence	1	2%
		intelligence	1	2%
Religion 2	Sujets	autre	22	28%
		sexualité	8	10%
		parole	1	1%
		dragué	1	1%
		nourriture	2	3%
		violence	19	24%
		origine	12	16%
		physique	3	4%
		éducation	2	3%
		2 sujets	1	1%
		maladie	1	1%
		sale	2	3%
		intelligence	1	1%
		infériorité	3	4%
Surpoids 1	Sujets	physique	26	38%
		autre	11	16%
		parole	11	16%
		sexualité	5	8%
		maladie	1	1%
		violence	13	19%
		intelligence	1	1%
		origine	1	1%

Surpoids 2	Sujets	physique	29	30%
		parole	21	21%
		jalousie	1	1%
		infériorité	4	4%
		autre	18	18%
		intelligence	15	15%
		2 sujets	1	1%
		violence	8	8%
		maladie	1	1%
		sexualité	1	1%
Surpoids 2a	Sujets	intelligence	10	19%
		sexualité	1	2%
		physique	17	32%
		autre	11	21%
		nourriture	2	4%
		parole	3	6%
		violence	2	4%
		2 sujets	1	2%
		infériorité	2	4%
		éducation	1	2%
		jalousie	2	4%

Grâce au tableau et aux pourcentages, nous pouvons voir que les deux conversations sur le thème de l'ethnicité ont des résultats différents, puisque dans la première, les sujets les plus récurrents sont “parole” avec 18%, “origine” avec 17% et “physique” avec 12%, tandis que dans la deuxième conversation, les sujets les plus récurrents sont “sexualité” avec 21%, “physique” avec 18% et “dragues” avec 12%. On peut donc voir que les sujets premiers de ce thème, c’est-à-dire “origine” et “physique” ne sont pas les plus récurrents et que la première conversation juge plus sur la parole des membres tandis que l’autre juge plutôt sur l’aspect sexuel, le couple.

Pour le thème de l’homophobie, on remarque également que ce ne sont pas exactement les mêmes sujets qui sont les plus abordés puisque “Homophobie 1” juge surtout la violence avec 27%, la parole avec 16% et la sexualité avec 10% et “Homophobie 2” comporte surtout les sujets “sale” et “orientation sexuelle” avec 24% et “sexualité” avec 17%. Donc, dans la première conversation, les auteurs associent l’homosexualité est surtout associée à la violence et jugent aussi la sexualité des personnes homosexuelles. Dans la deuxième conversation, les auteurs jugent les homosexuels comme étant “sales” et semblent les harceler pour leur orientation sexuelle, c’est-à-dire pour le fait qu’ils aiment quelqu’un du même sexe.

La plupart des sujets visés dans le thème “religion” sont la violence avec 24%, l’origine avec 16% et la sexualité avec 10% ce qui est étonnant dans ce thème puisque le sujet de la sexualité et tout ce qui est en rapport avec le sexe est généralement tabou dans la religion.

Dans les conversations sur le surpoids, on voit très clairement que le sujet le plus fréquent est “physique”. C’est le premier thème où toutes les conversations de celui-ci ont comme sujet le plus fréquent, celui le plus en rapport avec le thème. Donc, les membres ont bien respecté le scénario et principalement harcelé sur le physique.

3. Analyse du ton humoristique

Pour cette analyse, nous avons procédé de la même manière que pour celle des tweets, c’est-à-dire, que lorsqu’il y a un mot-clé lié à l’humour que ce soit “mdr”, “lol”, “ptdr”... nous comptons le message comme ayant un ton humoristique, seulement, nous ne pouvons pas analyser quel type d’humour il s’agit puisque nous n’avons pas d’indices comme #ironie, #humour. Ensuite, il nous a également suffit de compter et calculer les pourcentages des récurrences.

Thème	Analyse	Valeur	Récurrence	Pourcentage
Ethnicité 1	Ton humour	oui	8	9%
		non	86	91%
Ethnicité 2	Ton humour	oui	0	0%
		non	33	100%
Homophobie 1	Ton humour	oui	1	2%
		non	61	98%
Homophobie 2	Ton humour	oui	0	0%
		non	65	100%
Religion 2	Ton humour	oui	3	4%
		non	75	96%
Surpoids 1	Ton humour	oui	2	3%
		non	67	97%
Surpoids 2	Ton humour	oui	6	6%
		non	93	94%
Surpoids 2a	Ton humour	oui	2	4%
		non	50	96%

Grâce au tableau que nous obtenons après les calculs, nous pouvons remarquer qu’il y a peu d’humour dans ces conversations (< 10%) et qu’il y en a même qui n’en ont pas du tout (“ethnicité 2” et “homophobie 2”).

La conversation qui a le plus d’humour est “Ethnicité 1” avec 9% suivie de “Surpoids 2” avec 6%, puis “Religion 2” et “Surpoids 2a” avec 4%, “Surpoids 1” avec 3% “homophobie 1” avec 2% et “Ethnicité 2” puis “Homophobie 2” avec 0%.

On peut remarquer que le thème sur le surpoids est celui qui comporte le plus de ton humoristique (2 conversations surpoids parmis les 4 conversations ayant le plus d'humour), que les auteurs se permettent plus de plaisanter sur ce sujet qu'un autre ou alors qu'ils utilisent l'humour pour se cacher et ne pas révéler ce qu'ils pensent vraiment. Ils laissent alors planer le doute s'ils pensent vraiment ce qu'ils disent ou non. Ce qui est surprenant par rapport aux résultats du ton humoristique sur les tweets dans la partie 1, c'est que les tweets sur la religion comportaient peu voire pas du tout d'humour alors qu'ici, la conversation sur ce thème fait partie des trois conversations qui en comportent le plus. On peut donc penser que les auteurs de messages de haine sur Twitter ne font pas d'humour quand les messages sont publics mais qu'ils se le permettent en petit comité, lors d'une conversation de groupe, et donc qu'ils se permettent de plaisanter sur ce sujet.

4. Analyse des messages implicites

Pour réaliser cette analyse, nous avons relevé les messages pour lesquels l'insulte ou le message de haine est indirect, que ce soit des métaphores ou des sous-entendus. Par exemple dans l'intervention : "<David-Harceleur> espece de boule je te pousse tu roule ". "boule" + "tu roule" font une insulte implicite, c'est même une métaphore qui permet de juger la victime comme étant grosse et de la comparer à une boule au point même que la personne "roule". Les métaphores sont utilisées pour permettre de mieux visualiser de manière subtile, sans avoir recours à la comparaison explicite (comme, tel, semblable à…). De plus, elles sont utiles pour éviter de se répéter tout en conservant et en insistant sur un aspect de la caractérisation du sujet.

Une fois que nous avions repéré tous les messages indirects, nous les avons comptés puis réalisé une nouvelle fois des pourcentages. Ce qui a donné le tableau ci-dessous :

Thème	Analyse	Valeur	Récurrence	Pourcentage
Ethnicité 1	Implicite	oui	17	18%
		non	77	82%
Ethnicité 2	Implicite	oui	4	12%
		non	29	88%
Homophobie 1	Implicite	oui	9	15%
		non	53	85%
Homophobie 2	Implicite	oui	2	3%
		non	63	97%
Religion 2	Implicite	oui	3	4%
		non	75	96%
Surpoids 1	Implicite	oui	19	28%
		non	50	72%
Surpoids 2	Implicite	oui	9	9%
		non	90	91%
Surpoids 2a	Implicite	oui	4	8%
		non	48	92%

Le tableau nous permet de voir que “surpoids 1” est la conversation qui comporte le plus de messages “implicites” avec 28% suivie de “ethnicité 1” avec 18%, “homophobie 1” avec 15%, “ethnicité 2” avec 12% puis “surpoids 2” avec 9%, “surpoids 2a” avec 8%, “religion 2” avec 4% et enfin “homophobie 2” est la conversation comportant le moins de messages implicites avec 3%.

On remarque que les conversations sur le thème de l’ethnicité sont celles qui comportent le plus de messages indirects par rapport aux autres thèmes puisque les deux conversations de ce thème se trouvent dans les quatre conversations comportant le plus ce type de messages. Les auteurs de ce thème ont donc beaucoup usé des métaphores et des sous-entendus, qui, comme l’humour, permettent de laisser le choix à l’interlocuteur de la manière dont il veut comprendre le message. Et donc, cela leur permet de faire passer des messages sans dire explicitement les choses et donc, ils se cachent une nouvelle fois derrière leur discours, comme s’ils voulaient dire quelque chose sans vraiment l’assumer.

Cependant, pour les conversations “religion 2” et “homophobie 2”, les auteurs semblent au contraire assumer leurs paroles et ne passent pas par quatre chemins pour dire ce qu’ils pensent. Ce qui montre aussi qu’ils ne craignent pas les représailles ni les conséquences que peuvent engendrer leurs paroles.

Conclusion

Ce mémoire a pour ambition de relever de nouvelles données de messages de haine sur Twitter et lors d'une expérimentation, puis de les annoter et les analyser pour permettre à un algorithme supervisé de les repérer tout seul par la suite.

Il a fallu dans un premier temps définir le cyber-harcèlement, observer les travaux déjà réalisés, mettre en place une stratégie pour récolter ces nouvelles données puis les annoter et les analyser.

Nous avons pu réaliser ce travail minutieux grâce à l'analyse du discours mais aussi avec l'aide de la linguistique computationnelle. Ainsi, cela nous a permis de rendre compte des différentes cibles visées selon les thèmes abordés, des sujets qui sont la cause de ce harcèlement mais aussi de la manière dont les auteurs harcèlent : de manière directe, implicite avec des sous-entendus ou des métaphores ou alors, en utilisant l'humour.

Il nous a donc fallu du temps pour recueillir ces données que ce soient des tweets (messages de haine plus ponctuels et visant des groupes généraux) ou des messages au sein d'une conversation (qui sont plus insistants et visent quelqu'un en particulier) et les annoter avec précision, puisqu'une fois les annotations faites, la plus grosse partie était réalisée. Ce qui montre que former un algorithme va prendre du temps mais qu'une fois fait, cela vaudra le coup en terme d'efficacité pour repérer bien plus rapidement et précisément le message de haine que ce soit dans le discours, sur les plateformes en ligne et notamment lors de tweets postés sur Twitter, en allant au-delà de repérer uniquement les insultes.

Ce mémoire étant une première dans la constitution d'un corpus en français sur le cyber-harcèlement, il pourrait être intéressant de poursuivre les recherches pour obtenir encore plus de données et de pousser les analyses plus loin, en faisant analyser chaque aspect par un professionnel (exemple : que le ton humoristique soit analysé par un linguiste spécialiste de cette thématique) ou même de réaliser d'autres expérimentations.

Références bibliographiques

- Règles de Twitter
- Chiffres Internet Chiffres clés d'Internet et des réseaux sociaux en France en 2021
- Cyber-harcèlement (harcèlement sur internet) | service-public.fr.
- Chiffres Twitter - 2021
- Traitement automatique des langues
- Apprentissage supervisé
- Cyberviolence et cyberharcèlement: approches sociologiques
- Le cyberharcèlement chez les jeunes
- Chapitre 2. Le discours de haine
- Harcèlement entre pairs en milieu scolaire, quelle est l'ampleur de ce phénomène ?
- Chapitre 3. Discours de haine en ligne et réseaux sociaux
- Mik3M4n/help_dicrah: A ML-based hate speech detection tool on Twitter (in French)
- A Multilingual Evaluation for Online Hate Speech Detection
- OTESIA
- Creep: Home
- DELIVERABLE CREEP2
- HKUST-KnowComp/MLMA_hate_speech: Dataset and code of our EMNLP 2019 paper "Multilingual and Multi-Aspect Hate Speech Analysis"
- An Annotated Corpus for Sexism Detection in French Tweets
- L'humour dans les diverses formes du rire