Faculteit Letteren en Wijsbegeerte
Bachelorscriptie Taal- en Letterkunde
afstudeerrichting Bachelor Frans – Italiaans

La détection automatique de cas de cyberharcèlement textuel dans les médias sociaux

Maxim Baetens

Promotor: prof. dr. Walter Daelemans
Copromotor: prof. dr. Patrick Dendale
Assessor: Frederik Vaassen

Universiteit Antwerpen
Academiejaar 2012 – 2013

Faculté Lettres & Philosophie
Mémoire de licence : Lettres et Langues
option français – italien

La détection automatique de cas de cyberharcèlement textuel dans les médias sociaux

Maxim Baetens

Directeur de mémoire: prof. dr. Walter Daelemans
Codirecteur de mémoire: prof. dr. Patrick Dendale
Assesseur: Frederik Vaassen

Université d'Anvers
Année universitaire 2012 – 2013

Le soussigné, Maxim Baetens, étudiant en Lettres et Langues (français – italien), déclare que ce mémoire de licence est tout à fait original et qu'il a été rédigé uniquement par lui-même. Pour toutes les informations et idées empruntées, le soussigné a indiqué explicitement et en détail les sources consultées.

Anvers, le 5 mai 2013

Signature

Table des matières

0. Introduction

En septembre 2011, Jamey Rodemeyer[1], un garçon de quatorze ans, s'est pendu après avoir été l'objet, pendant des années, de (cyber)harcèlement à cause de son orientation sexuelle. En octobre 2012, la jeune Canadienne Amanda Todd[2] se suicide suite à des brimades, physiques et en ligne, constantes. En janvier 2013, des individus anonymes ont créé une page Facebook portant le nom « *Antwerpse hoeren* » (putes anversoises) où les utilisateurs du réseau social pouvaient commenter, sans se retenir, des photos de filles innocentes.

Les tragédies mentionnées ci-dessus sont loin d'être des cas uniques. Depuis l'introduction d'Internet, toutes sortes de nouveaux phénomènes sociologiques ont vu le jour. Une des évolutions les plus troublantes est le cyberharcèlement (le harcèlement en ligne). Les dernières années, on en entend parler davantage dans les médias, surtout quand la situation a mal tourné. Les experts essayent de sensibiliser la population à faire attention aux dangers en ligne. Malheureusement, les actions des spécialistes n'ont pas d'influence durable : certaines gens continuent à causer des ennuis à autrui. Le cyberharcèlement est d'autant plus menaçant, qu'Internet de nos jours est présent partout et toujours (pensons entre autres aux ordiphones, aux portables et – plus récemment – aux tablettes tactiles).

Ce mémoire se veut d'offrir un aperçu des techniques informatiques possibles pour détecter automatiquement des cas de cyberharcèlement textuel dans les médias sociaux. Nous définissons un *cas de cyberharcèlement* comme « une conversation en ligne qui se déroule entre au moins deux utilisateurs d'un même réseau social et qui contient au minimum un message (ou commentaire) dépréciatif. » Notre objectif est de créer un système (semi-)automatique qui aide les modérateurs des réseaux sociaux à sélectionner les cas de cyberharcèlement graves et à intervenir adéquatement (en l'occurrence, en supprimant le message blessant). Ce système peut être composé de plusieurs algorithmes, de manière à ce que la détection soit plus précise. De telle façon, nous voulons contribuer à créer un monde virtuel sûr pour tout le monde et aider à éviter que d'autres situations de cyberharcèlement se terminent en drames.

Les questions que nous nous poserons seront les suivantes :

1) Quels traits (linguistiques) caractérisent les messages des victimes de cyberharcèlement et des cyberharceleurs dans les médias sociaux?
2) Quelles techniques la linguistique informatique offre-t-elle pour détecter les traits dégagés des messages de cyberharcèlement ?
3) Comment peut-on éviter que le cyberharcèlement se poursuive après la détection automatique?

Notre travail commencera par l'étude de la notion de *cyberharcèlement*. Nous confronterons le harcèlement dit « classique » à sa variante moderne, le cyberharcèlement, pour mieux comprendre la portée de la problématique. Nous nous arrêterons également sur les solutions sociologiques, pédagogiques et - le cas

[1] Praetorius, Dean, 'Jamey Rodemeyer, 14-year old boy, commits suicide after gay bullying, parents carry on message', *The Huffington Post*, le 22 septembre 2011. [disponible en ligne:] http://www.huffingtonpost.com/2011/09/20/jamey-rodemeyer-suicide-gay-bullying_n_972023.html [30/01/2013].
[2] Grenoble, Ryan, 'Amanda Todd : bullied Canadian teen commits suicide after prolonged battle online and in school', *The Huffington Post*, le 11 octobre 2012. [disponible en ligne:] http://www.huffingtonpost.com/2012/10/11/amanda-todd-suicide-bullying_n_1959909.html [30/01/2013].

échéant - informatiques qui sont disponibles à l'heure actuelle pour réduire le nombre de cas de cyberharcèlement. Nous esquisserons un profil du cyberharceleur et de la victime typiques dans le but de découvrir les traits (linguistiques) que les techniques computationnelles peuvent détecter.

Dans le deuxième chapitre nous entrerons dans les détails de la détection de cas de cyberharcèlement. Nous analyserons d'abord un exemple d'un cas de cyberharcèlement pour en dégager les traits qui caractérisent le phénomène. Nous présenterons ensuite les différentes techniques et méthodes informatiques permettant de trouver automatiquement des cas de cyberharcèlement à partir de ces traits. Nous examinerons l'efficacité des techniques ainsi que les difficultés qu'elles connaissent.

Dans le troisième chapitre, nous commenterons quelques modes d'intervention possibles après la détection d'un cas de cyberharcèlement. Les modérateurs peuvent par exemple supprimer simplement un message blessant, mais de telle manière l'auteur du commentaire vexatoire ne prendra pas conscience de ce qu'il était sur le point de déclencher. Nous préférons que le système intervienne de manière efficace au moment nécessaire. Dès que le système informatique détecte un cas de cyberharcèlement, le réseau social peut intervenir directement. Nous formulerons quelques idées sur une intervention adéquate.

Le dernier chapitre présentera nos conclusions et nos réponses aux questions de recherche. Nous développerons aussi des pistes de recherche à suivre dans l'avenir.

Avant d'entrer en matière, nous tenons à remercier les personnes sans qui ce mémoire n'aurait pas été possible. Premièrement, le directeur de mémoire, le prof. dr. Walter Daelemans, qui a accepté notre sujet et qui nous a toujours donné son avis franc sur le texte. En cas de doute, c'était la première personne à qui s'adresser. Deuxièmement, le codirecteur de mémoire, le prof. dr. Patrick Dendale, qui nous a averti quand il y avait de gros problèmes de langue et qui a pris le temps de nous expliquer patiemment les erreurs commises. Sans son aide, le mémoire aurait été beaucoup moins clair. Et finalement, nous remercions nos parents qui nous ont toujours encouragé au cours de notre formation.

1. Qu'est-ce que le cyberharcèlement?

Contrairement au harcèlement classique, le cyberharcèlement est un phénomène plutôt récent. Il est fort répandu dans les écoles : en 2006, 61,9 % des élèves flamands interrogés ont déclaré avoir été victimes d'un cas de cyberharcèlement (Vandebosch e.a., 2006 : 177) [3]. Quoique à peu près deux tiers des écoliers aient déjà eu affaire avec ce problème, il reste difficile de bien le décrire et délimiter. Dans ce chapitre nous examinerons en quoi le harcèlement classique diffère du cyberharcèlement et quels sont les traits communs. Nous regarderons de plus près le profil des cyberharceleurs et des victimes pour voir s'il y a des tendances générales à remarquer. Nous focaliserons sur les caractéristiques qui nous semblent pertinentes pour la détection automatique du cyberharcèlement (l'âge, le sexe et l'émotion); nous ne visons donc pas à esquisser un profil exhaustif.

1.1 Le harcèlement classique *versus* le cyberharcèlement

Tous les cas de harcèlement classique ont un déroulement similaire et des rôles plus ou moins récurrents. Greene (2000) nomme cinq caractéristiques qui définissent un cas de harcèlement classique: (1) le harceleur a l'intention de faire peur à sa victime ou de la maltraiter (2) à plusieurs reprises ; (3) la victime fait partie du groupe social du harceleur ; (4) elle n'a pas provoqué le harceleur ni (non-)verbalement ni physiquement; (5) ils entretiennent une relation de force déséquilibrée, c'est-à-dire le harceleur se sent plus puissant que la victime.

Deux personnes au moins sont impliquées : une victime et un harceleur. Éventuellement quelques témoins sont sur la touche. Ceux-ci adoptent souvent le rôle de spectateur, qui n'agit pas de peur qu'il devienne la prochaine victime du harceleur (Harris & Petrie, 2002 ; Campbell, 2005). Salmivalli (1999 : 454) distingue encore plusieurs autres rôles apparaissant lors d'un harcèlement : « l'assistant » du harceleur, « le renforçateur », qui encourage le harceleur à continuer (entre autres en approuvant le comportement de celui-ci) et « le défenseur » qui vient en aide à la victime.

Le harceleur peut employer des techniques physiques ou verbales pour importuner sa victime. Par harcèlement physique on entend frapper, pousser, battre ou effectuer d'autres actes violents. La destruction d'objets personnels de la victime (par exemple son cartable) appartient également à cette catégorie. Le harcèlement verbal, en revanche, consiste à vexer ou à menacer quelqu'un au moyen d'injures et d'insultes. Finalement, existe encore un harcèlement non-verbal qui se réalise par des gestes obscènes (par exemple le doigt d'honneur) et l'exclusion d'un groupe social. (Vandebosch & Van Cleemput, 2009 ; voir aussi la grille 1).

Après ce passage sur le harcèlement classique, nous comparons le harcèlement traditionnel avec sa variante plus moderne. La description du cyberharcèlement part souvent de celle du harcèlement classique, sauf qu'on y ajoute l'usage de moyens technologiques (tels qu'un portable, Internet et des SMS). Comme l'ont constaté Vandebosch & Van Cleemput (2009), il n'est pas facile de comparer les deux types de harcèlement. Il est vrai que d'ordinaire la victime ne provoque pas le comportement du cyberharceleur et que celui-ci a toujours l'intention de blesser sa victime. Avec le harcèlement classique, il est facile de comprendre l'intention du

[3] Les chercheurs ont effectué un sondage auprès de 1416 jeunes, entre 10 et 18 ans. Ils considéraient comme cyberharcèlement tous les actes blessants faits par des moyens technologiques (SMS, Internet, MMS,…). Les jeunes devaient indiquer s'ils avaient vécu quelques situations formulées. Il faut noter que les situations que les chercheurs ont proposées ne sont pas nécessairement considérées comme des cas de cyberharcèlement par les jeunes – il se peut que ceux-ci aient une autre acception de ce que c'est que le cyberharcèlement.

harceleur (par exemple, un harceleur qui frappe veut clairement blesser sa victime) Le cyberharcèlement, par contre, n'est pas toujours tellement clair. Il faut tenir compte du fait qu'un message électronique peut facilement être mal compris. Ainsi, en l'absence de langage du corps, une mauvaise interprétation d'un message innocent pourrait donner lieu à une aggravation d'une situation normale.

Les nouveaux moyens technologiques nous forcent à réinterpréter deux critères de Greene: la répétition et l'inégalité de pouvoir. Avec le cyberharcèlement, ils ont tous deux une acception plus nuancée.

Premièrement, il est important que le comportement soit intentionnellement répétitif: le cyberharceleur répète sans arrêt ses injures pour blesser sa victime. C'est la répétition qui distingue le harcèlement du taquinage: un ami, par exemple, pourrait parfaitement pour taquiner poster un message (considéré aussi comme offensif) sur le mur d'un réseau social d'un copain sans avoir l'intention de lui faire mal (Langos, 2012).

En outre, le caractère répétitif provient aussi du fait qu'un message public reste longtemps consultable et qu'un nouvel utilisateur qui lit ce message 'répète' en quelque sorte le harcèlement. Slonje & Smith (2007 : 154) signalent que « chaque clic pourrait compter pour une répétition », car il y a toujours un aspect sensible de la personnalité de la victime qui est rendu public aux visiteurs (parfois anonymes) de sa page personnelle.

Deuxièmement, l'inégalité du pouvoir pose problème. À part les forums où les modérateurs contrôlent les messages publiés, on ne trouve pas de vraie hiérarchie fixe sur Internet. Ybarra & Mitchell (2004 : 1313) soutiennent que le cyberharceleur peut maintenir son pouvoir en cachant son identité (*masquerading*). Il crée un faux compte derrière lequel il se cache pour dire tout ce qu'il veut ; c'est que personne ne connaît sa véritable identité. Li (2006 : 158) remarque que « les gens ont l'impression qu'Internet est impersonnelle et qu'ils peuvent par conséquent dire n'importe quoi ». Il n'est même pas exclu que la victime ne connaisse pas personnellement son cyberharceleur ; le cyberharcèlement ne se limite donc pas au même groupe social.

Sur l'effet de l'anonymat dans les communautés sociales, Omernick & Sood (2013) ont effectué une recherche. Ils partent d'une notion introduite en psychologie, à savoir la « dé-individuation ». Quand une personne est placée dans un groupe (le groupe d'internautes), elle ne se profile plus comme individu ; elle fait partie du groupe homogène sans qu'elle se distingue. Cette perte d'identité crée une « dissociation de la conscience de soi », d'où vient la diminution du contrôle de soi. Voilà pourquoi le cyberharceleur terrorise sa victime sans pitié dans l'anonymat. En plus, faute de réaction immédiate de la victime, le cyberharceleur ne ressent plus de restrictions. Il se croit libre et il fait ce qu'il veut, quand et avec qui il le veut.

Omernick & Sood (2013) ont pris un corpus dans *Techcrunch.com*, un site internet qui a remplacé en mars 2011 sa messagerie avec l'option 'anonyme' par la plateforme de *Facebook*. Ils voulaient voir quelles étaient les conséquences, si les utilisateurs ne pouvaient pas commenter anonymement. Qualitativement parlant, ils ont constaté que les messages liés à un compte *Facebook* étaient d'une meilleure qualité (avec moins d'injures et moins d'émotions négatives) par rapport aux messages anonymes. Quantitativement parlant, les chercheurs n'ont pas vu de différence en ce qui concerne la participation. Au contraire, les utilisateurs identifiés participaient plus et plus longtemps à une discussion. Selon Omernick & Sood (2013), l'amélioration de la qualité et le manque de baisse de la participation sont des arguments en faveur d'une communauté qui n'offre pas la possibilité de poster des messages anonymes. Ils avertissent toutefois qu'il faut tenir compte de la nature spécifique de chaque communauté et des répercussions légales.

Même si les utilisateurs opèrent dans l'anonymat, on distingue toujours les rôles du harcèlement classique dans les cas de cyberharcèlement. Aux six rôles déjà mentionnés, Xu e.a. (2012b) ajoutent un reporter et un accusateur. Le reporter a été témoin d'un cas de cyberharcèlement et il rapporte ce qu'il a vu. L'accusateur accuse directement un cyberharceleur de son comportement inapproprié. Les chercheurs distinguent donc huit rôles dans chaque cas de cyberharcèlement. Évidemment, il se peut qu'une seule personne assume plusieurs rôles. Nous reprenons ici la figure de Xu e.a. (2012b) dans laquelle il devient clair quels rapports entretiennent les divers rôles.

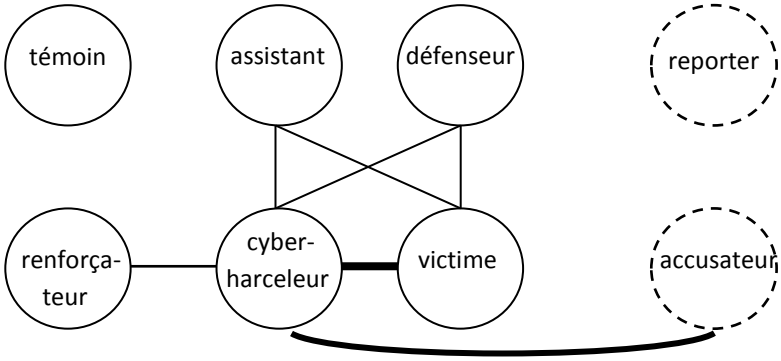

Figure 1: Les rôles dans un cas de cyberharcèlement (Xu e.a., 2012b : 657) [traduction]
(Les lignes grasses indiquent un lien actif et direct ; les pointillés sont les rôles spécifiques du cyberharcèlement.)

Il est important aussi de souligner que le cyberharcèlement a une portée large. Alors que le harcèlement classique se limite principalement à la cour de l'école ou aux organisations de jeunesse, le cyberharcèlement ne connaît pas ces limites. Il n'est pas exclu qu'une victime soit harcelée aussi bien à l'école qu'à la maison, de façon qu'elle n'échappe jamais à ses agresseurs. Ou bien c'est l'inverse qui se produit : une victime agressée a la possibilité de se défouler en cherchant à son tour une victime et en devenant ainsi harceleuse elle-même. La même chose vaut pour le harceleur : bien que soutenu par ses camarades à l'école, il se retrouve seul sur Internet. Par conséquent, il est possible qu'il devienne victime dans le monde virtuel (Ybarra & Mitchell, 2004 ; Vandebosch e.a ., 2006).

Les divers moyens technologiques permettent une variété de manières pour importuner une victime[4]. Il va de soi qu'un cyberharceleur ne peut pas blesser physiquement sa victime. Mais un cyberharceleur peut détruire des biens informatiques à distance, plus particulièrement en envoyant un virus informatique ou en piratant l'ordinateur ou le portable de quelqu'un d'autre. Ce sont surtout les cas de cyberharcèlement (non-)verbaux qui sont beaucoup plus fréquents par rapport au harcèlement classique : pensons aux commentaires offensifs postés arbitrairement sur le compte d'une victime (le soi-disant *flaming*) ou aux photos retouchées. Sur le plan social, on peut exclure quelqu'un d'un groupe en ligne, pour lui empêcher de parler avec les autres.

Alors que le harcèlement classique se déroule surtout directement, le cyberharcèlement offre plusieurs possibilités pour harceler une victime de façon indirecte. Un comportement indirect signifie que le

[4] La plupart des exemples ont été pris dans Vandebosch & Van Cleemput (2009)

cyberharceleur ne dit pas ses mots offensifs en face de la victime, mais qu'il les écrit en son absence. Le cyberharceleur peut par exemple envoyer un courriel à tous ses contacts dans lequel il raconte des histoires personnelles et/ou humiliantes sur la victime. Dans la grille 1 nous donnons un aperçu des types les plus fréquents de (cyber)harcèlement.

	le harcèlement classique	le cyberharcèlement
direct	- physique (p.e. frapper quelqu'un) - propriété (p.e. endommager les biens de quelqu'un) - verbal (p.e. insulter quelqu'un) - non-verbal (p.e. faire des gestes obscènes) - social (p.e. exclure quelqu'un d'un groupe)	- propriété (p.e. envoyer un virus informatique) - verbal (p.e. *flaming*) - non-verbal (p.e. retoucher des photos) - social (p.e. exclure quelqu'un d'une communauté en ligne)
indirect	- propager des rumeurs	- publier un courriel contenant des informations confidentielles - assumer une autre identité (*masquerading*) pour poster des messages blessants - faire des ragots sur quelqu'un - organiser des sondages en ligne sur quelqu'un - créer un faux compte

Grille 1: les divers types de (cyber)harcèlement (Vandebosch & Van Cleemput, 2009: 1353) [traduction]

1.2 Profil du cyberharceleur

Notre étude générale a montré qu'au moins deux rôles font nécessairement partie d'un cas de cyberharcèlement : le cyberharceleur et la victime. Les deux parties suivantes proposent un examen des profils typiques des deux rôles. Nous commencerons par composer un profil du cyberharceleur stéréotypé, mais nous ne traiterons que les caractéristiques intéressantes pour la linguistique informatique, à savoir l'âge et le sexe (nous ne traiterons donc pas du niveau d'instruction, du lieu de résidence ni d'autres facteurs sociologiques).

Certains chercheurs voient un rapport direct entre le harcèlement classique et le cyberharcèlement: un harceleur peut facilement devenir cyberharceleur (Slonje & Smith, 2007; Vandebosch e.a., 2006 ; Vandebosch & Van Cleemput, 2009 ; Ybarra & Mitchell, 2004). Cela n'implique pas forcément que tous les cyberharceleurs soient également des harceleurs dans la vie réelle. Plusieurs scientifiques considèrent que les risques principaux pour devenir cyberharceleur sont : 1) des parents moins contrôleurs et 2) de bonnes connaissances d'Internet et des moyens technologiques tout court. Langos (2012 : 287) soulève la possibilité qu'une victime puisse se sentir impuissante face aux connaissances technologiques supérieures d'un cyberharceleur. Pour cette raison, l'inégalité du pouvoir est de nouveau renforcée : la victime n'échappe jamais au cyberharceleur, puisqu'il trouve toujours des moyens pour atteindre sa victime grâce à ses connaissances technologiques.

Quant au sexe, Li (2006 : 8) et Vandebosch e.a. (2006 : 135) signalent que ce sont le plus souvent les garçons qui ont recours au cyberharcèlement. Mais dans une enquête que Vandebosch & Van Cleemput (2009 : 1365) ont effectuée plus tard, elles constatent qu'il y a autant de cyberharceleurs que de cyberharceleuses. Selon elles, le sexe ne serait donc pas un facteur déterminant.

En ce qui concerne l'âge, Vandebosch e.a. (2006 : 136) ont trouvé que les cas de cyberharcèlement dépendent de l'âge, initiant dès que les enfants ont un accès libre aux moyens technologiques. Plus les jeunes

avancent dans l'âge, plus ils ont affaire au cyberharcèlement. Le nombre de cas de cyberharcèlement continue à augmenter jusqu'à un certain âge, après lequel on constate une baisse (à partir de 20 ans). Li (2006 : 10) remarque que l'adolescence (qui va grosso modo de dix à vingt ans[5]) est une « période de brutalités » où l'on tente de développer sa propre identité. Les individus qui ne suivent pas les normes sociales courantes risquent de devenir victimes d'un cyberharceleur. Après l'adolescence on tient moins à établir son identité, puisqu'on se l'est déjà plus ou moins formée.

1.3 Profil de la victime

Tout comme nous l'avons fait pour le cyberharceleur, nous décrirons la victime typique. Vandebosch e.a. (2006 : 181) voient une corrélation entre les victimes du cyberharcèlement et celles du harcèlement classique. Qui est importuné à l'école a plus de risques de devenir aussi victime d'un cyberharceleur. Les victimes des deux types de harcèlement ont beaucoup de caractéristiques en commun, sauf que les victimes du cyberharcèlement vivent encore plus dans une peur absolue. Elles ne doivent pas seulement avoir peur de recevoir un message d'un cyberharceleur, mais aussi de recevoir des réactions d'autres personnes concernées. Cette peur continue fait que les victimes souffrent de maladies mentales: elles ont dans la plupart des cas une vision sombre, négative, voire suicidaire, du monde.

Comme c'est le cas pour les cyberharceleurs, les scientifiques ne se mettent pas d'accord sur le rôle que joue le sexe. Vandebosch e.a . (2006 : 181) posent que la majorité des victimes d'un cyberharceleur sont des filles. Li (2006 : 7) et Ybarra & Mitchell (2004 : 1309) par contre ne constatent pas de différences significatives entre le nombre de victimes masculines et féminines.

Suivant les caractéristiques du cyberharceleur, la majorité des victimes ont atteint l'âge (pré)adolescent. Vu que les adolescents emploient les moyens technologiques le plus activement (Vandebosch e.a., 2006 : 136), il n'est pas étonnant que nous retrouvions pas mal de cyberharceleurs et de victimes dans le même groupe d'âge. Ce qui n'exclut évidemment pas que des personnes plus âgées ne puissent pas être touchées du cyberharcèlement.

1.4 Solutions

Avec le harcèlement classique, l'école peut encore prendre des mesures afin d'éviter une répétition du comportement négatif. Le cyberharcèlement en revanche peut continuer en dehors des murs de l'école, s'introduisant même dans la maison de la victime. Il n'est pas surprenant que l'intervention dans l'ambiance personnelle des élèves pose plus de problèmes à la direction. Campbell (2005 : 72) se pose la question de savoir si l'école a même droit d'interdire l'emploi des moyens technologiques à la maison. Strictement parlant, ce sont les parents qui sont responsables des actes de leurs enfants chez eux.

Un deuxième problème qui pose des problèmes est celui de l'anonymat. Étant donné qu'un cyberharceleur peut adopter un pseudonyme, il est souvent compliqué de trouver la vraie identité de l'auteur (Campbell, 2005 : 72). En plus, un seul auteur peut poster sous plusieurs pseudonymes ou plusieurs auteurs peuvent

[5] L'OMS définit *adolescent* comme « un individu qui a entre dix et dix-neuf ans » et *jeune* comme « un individu qui a entre quinze et vingt-quatre ans.» [traduction] (WHO, *Child and Adolescent Health*. http://www.who.int/features/factfiles/adolescent_health/en/index.html [20/03/2013].

adopter le même nom d'utilisateur (Omernick & Sood, 2013). D'après Li (2006 : 3), la liberté d'expression fait qu'il est extrêmement difficile de supprimer un seul message ou même un site internet complet, encourageant ainsi le caractère répétitif du cyberharcèlement.

Vu que l'école ne peut pas intervenir, les parents doivent mieux contrôler et informer leurs enfants. Il est nécessaire de montrer aux enfants les conséquences du cyberharcèlement. Pour y parvenir, les parents doivent d'abord connaître eux-mêmes les dangers d'Internet (y compris le cyberharcèlement). En outre, les adultes doivent se renseigner sur les nouvelles technologies. Les enfants d'aujourd'hui ont grandi avec l'ordinateur et souvent ils ont des compétences technologiques plus développées par rapport à leurs parents (Campbell, 2005 : 71). Un père ou une mère peuvent avoir des difficultés à reconnaître un cas de cyberharcèlement, parce qu'ils ne connaissent rien au phénomène. Si les instructeurs et les parents collaboraient plus étroitement, le nombre de victimes pourrait déjà diminuer. Les instructeurs spécialisés peuvent combler les lacunes en illustrant les risques à l'aide d'exemples réels. Comme le dit Chisholm (2006: 82):

> «The irony is that parents, sitting right next to their child who is in a dangerous situation online, may be totally clueless about the present danger because they and their child are in the comfort and safety of home. The parents, as a result, perceive no danger and therefore fail to protect their child. »

Puis, l'attitude des jeunes devrait changer. La moitié des victimes n'ont pas informé leurs parents de leur problème (Vandebosch e.a. , 2006 : 181). Campbell (2005) pense que c'est parce que les jeunes ont peur que leurs parents ne confisquent leurs moyens technologiques. Pour aboutir à une prévention efficace, il est nécessaire que la communication entre parents et jeunes soit plus transparente.

Il est possible que les parents doivent prendre des mesures qui vont plus loin de la prévention, comme l'emploi des logiciels de contrôle parental. Il existe des logiciels qui permettent aux parents de gérer le comportement en ligne de leur(s) enfant(s). Kontostathis e.a. (2010 : 11) donnent un aperçu des logiciels les plus renommés aux États-Unis. Nous reprenons ici quelques outils, tenant compte des éventuelles améliorations au fil des années. Premièrement, *Net Nanny*[6] enregistre les conversations de l'enfant dans les réseaux sociaux les plus populaires. En outre, le logiciel offre la possibilité de cacher les injures présentes dans les commentaires d'autres personnes. Deuxièmement, *Kidswatch*[7] fonctionne comme *Net Nanny*, mais il surveille aussi les messages qu'envoient les enfants. Les parents choisissent dans des listes d'injures (ou ils optent pour une liste rédigée par eux-mêmes) et *Kidswatch* les prévient par courriel lorsqu'il en détecte une dans un message. Troisièmement, *Bsecure*[8] avertit les parents quand il rencontre « une activité suspecte » dans une conversation en ligne de l'enfant. *Bsecure* prétend avoir accès à plus de 75 réseaux sociaux, garantissant ainsi un contrôle approfondi. Pour finir, *Iambigbrother*[9] est un logiciel espion (invisible) qui enregistre tout ce que l'enfant fait sur l'ordinateur : les sites internet qu'il visite, ses mots de passe et ses frappes lors d'une conversation de chat. En plus, il donne l'occasion de prendre une capture d'écran au moment où il détecte l'emploi d'une injure, d'un juron ou d'un terme marqué comme offensif. Il n'offre cependant pas de

[6] Net Nanny. http://www.netnanny.com [17/09/2012].
[7] Kidswatch. http://www.kidswatch.com [17/09/2012].
[8] Bsecure. http://www.bsecure.com [17 /09/2012].
[9] Iambigbrother. http://www.iambigbrother.com [17/09/2012].

surveillance dans les médias sociaux. Tous les logiciels cherchent surtout des injures à partir d'une liste préalablement rédigée ; ils ne tiennent pas compte du contexte, ni des traits linguistiques. À cela s'ajoute que les parents doivent prendre leurs responsabilités : pour bien estimer la gravité de la transgression ils doivent (de nouveau) se renseigner sur les dangers du cyberharcèlement.

Quant aux mesures des médias sociaux, ils engagent des modérateurs pour contrôler tous les messages qui paraissent sur un réseau social. Un réseau social permet aux utilisateurs de communiquer avec leurs amis et de partager leurs photos ou vidéos. Les amis ont également la possibilité de commenter une photo ou un message qui ont été postés. Plus le nombre d'inscriptions augmente, plus la tâche du modérateur devient compliquée. Voilà où se situe l'utilité de la détection automatique de cas de cyberharcèlement.

Même si les médias sociaux sont omniprésents, on n'a pas encore développé de logiciel qui facilite le travail des modérateurs (Ptaszynski e.a., 2010). Les pages 'Aides' des réseaux sociaux les plus connus (*MySpace, YouTube, Facebook, Twitter* et *Formspring.me*) donnent tous le même conseil au cas où l'on aurait affaire avec un cas de cyberharcèlement: la victime doit rédiger elle-même un rapport pour signaler un message de cyberharcèlement et entre-temps elle est conseillée de bloquer le compte du cyberharceleur. Notons toutefois que les consignes ne mentionnent pas ce que le réseau social comprend exactement par *cyberharcèlement*, on écrit seulement que les messages violant les conditions d'utilisation seront supprimés. Seuls *Formspring.me*[10] et *YouTube*[11] élaborent davantage la problématique: ils renvoient à des organisations d'aide (telles que les lignes d'aide) et ils suggèrent aussi de contacter une personne de confiance. Cependant, peu d'utilisateurs font emploi des directives offertes par les réseaux sociaux. Dans le deuxième chapitre de notre mémoire, nous parlerons plus en détail des différentes techniques informatiques pour construire un système semi-automatique qui détecte des messages de cyberharcèlement textuel.

1.5 Conclusions

Nous avons vu que le cyberharcèlement a des traits typiques qui diffèrent du harcèlement classique. En premier lieu, il y a le fait que le cyberharceleur peut opérer dans l'anonymat : il n'est pas obligé de révéler son identité. De telle façon, il augmente son pouvoir sur la victime. En deuxième lieu, le cyberharcèlement se produit souvent sur Internet, un endroit public d'accès facile. La portée est donc beaucoup plus étendue qu'elle ne l'est avec le harcèlement classique : l'agression continue loin des murs de l'école. En troisième lieu, il faut tenir compte des rôles : le cyberharcèlement ajoute deux rôles au six rôles du harcèlement classique (cyberharceleur, victime, témoin, renforçateur, assistant et défenseur) , à savoir le reporter et l'accusateur.

Nous avons constaté qu'il est difficile de créer un profil typique d'un cyberharceleur et d'une victime. Ni l'âge, ni le sexe des cyberharceleurs ou des victimes ne constituent un facteur déterminant pour prédire si quelqu'un court le risque de devenir victime ou cyberharceleur. Il paraît que le cyberharcèlement soit le plus fréquent pendant l'adolescence, mais cela ne signifie pas que des personnes d'autres tranches d'âge ne soient pas atteintes du phénomène. Les cyberharceleurs sont généralement moins contrôlés par leurs parents et ont

[10] Cyberbullying and Impersonation. https://formspringme.zendesk.com/entries/20923496-cyberbullying-and-impersonation [22/01/2013].
[11] Harcèlement et cyberintimidation. http://support.google.com/youtube/bin/answer.py?hl=fr&hlrm=nl&answer=126266&topic=2803240&ctx=topic [22/01/2013].

de bonnes connaissances quant à l'usage des moyens technologiques. Des victimes, on peut dire qu'elles sont souvent pessimistes, à cause de leur peur continue d'être cyberhacelées.

Étant donné que le cyberharcèlement ne se limite pas à la cour de récréation, il est difficile pour l'école d'intervenir. Les parents devraient donc s'occuper plus du comportement en ligne de leurs enfants, mais souvent ils sont ignorants dans l'utilisation des nouvelles technologies. Il s'agit de bien les informer sur le cyberharcèlement. Dans le domaine des réseaux sociaux, il est souhaitable que les modérateurs surveillent les messages postés pour intervenir le plus vite possible. Cependant, le nombre croissant d'utilisateurs rend ce travail impossible. Peut-être que la détection automatique peut alléger la tâche des modérateurs en les aidant à détecter les messages suspects. Regardons de plus près quelles techniques en linguistique informatique permettent la détection automatique de cas de cyberharcèlement textuel.

2. La détection automatique de cas de cyberharcèlement textuel dans les médias sociaux

Maintenant que nous avons décrit ce que c'est que le cyberharcèlement, passons à la détection computationnelle de messages de cyberharcèlement. Si les études sociologiques sur les causes et les effets du cyberharcèlement sont très nombreuses, peu de chercheurs se sont occupés du traitement automatique de ce problème fort répandu dans les médias sociaux. On peut se demander pourquoi nous examinons la possibilité d'automatiser la détection, vu que presque toutes les communautés sociales offrent la possibilité de signaler les commentaires inappropriés (c'est ce qu'on appelle *flagging*). Malgré le grand nombre d'utilisateurs des réseaux sociaux, il y en a peu qui font usage de cette fonction. En outre, Sood, Churchill & Antin (2012b) remarquent que certains utilisateurs signalent parfois un message pour la seule raison qu'il n'exprime pas leur propre opinion; le modérateur n'est donc pas sûr que les cas signalés soient nuisibles à autrui.

Ce chapitre se construira autour de deux questions: 1) Quels traits linguistiques ou métalinguistiques permettent de distinguer messages de cyberharcèlement et messages inoffensifs ? ; 2) De quelles techniques la linguistique informatique dispose-t-elle pour détecter ces traits et sont-elles efficaces? Afin de répondre à ces deux questions, nous présenterons d'abord une analyse détaillée d'une conversation où l'on retrouve clairement des traces de cyberharcèlement, et ensuite les techniques informatiques que nous retenons comme intéressantes pour la détection de cas de cyberharcèlement.

2.1 Quels traits linguistiques caractérisent le cyberharcèlement textuel?

Afin de pouvoir développer un système qui détecte des cas de cyberharcèlement, il est indispensable de dégager les caractéristiques du langage d'un cas de cyberharcèlement. Il faudra aussi résoudre un problème persistant qui complique la tâche des chercheurs, à savoir l'absence d'un corpus large et standard.

Quoique le cyberharcèlement soit un problème actuel et urgent, peu de chercheurs s'en sont occupés. Le manque d'un corpus standard et annoté est une des principales raisons pour lesquelles les chercheurs n'avancent pas beaucoup dans leurs recherches (Yin e.a., 2009). Bayzick, Kontostathis & Edwards (2011)[12], Xu e.a. (2012)[13] et CAW 2.0 (2009)[14] sont les seuls à avoir créé un corpus annoté, rendu disponible à tout le monde. Les entrées des corpus sont des messages complets, pris dans un réseau social spécifique et accompagnés d'informations supplémentaires sur leurs auteurs (le sexe, l'âge, le lieu de résidence,…).

Chacun de ces corpus contient des messages d'un seul et même réseau social ; un corpus qui serait un mélange d'énoncés collectionnés dans plusieurs réseaux aiderait peut-être à améliorer les techniques de détection de cas de cyberharcèlement. Si les chercheurs ont choisi de travailler sur un réseau en particulier, c'est parce que chaque communauté a d'autres normes et d'autres expressions (Sood, Antin & Churchill, 2012a). Ceci implique également que les systèmes de détection ne sont valables que pour un seul site internet. En rassemblant des données d'un grand nombre de réseaux, le système reconnaîtra plus facilement les cas de cyberharcèlement dans toutes leurs formes. Il se peut que l'auteur d'un commentaire fasse partie d'une

[12] MySpace Group Data Labeled for Cyberbullying. http://www.chatcoder.com/DataDownload [1/10/2012].
[13] Bullying Traces Data Set V1.0. http://research.cs.wisc.edu/bullying/data.html [13/10/2012].
[14] Fundación Barcelona Media Training Dataset. http://caw2.barcelonamedia.org/node/7 [28/10/2012].

certaine communauté sociale, mais écrive une seule fois un message blessant sur un autre réseau social en empruntant les normes à sa communauté préférée. Si les chercheurs prenaient en considération tous les médias sociaux, ils aboutiraient à un système qui détecterait mieux les différentes expressions propres au cyberharcèlement textuel. En même temps, le corpus serait étendu considérablement avec des données exemplaires de cyberharcèlement : n'oublions pas que parmi les centaines de milliers de messages qui sont postés quotidiennement dans les médias sociaux, il n'y a qu'un pourcentage fort limité d'énoncés dépréciatifs.

Pour ne pas perdre de temps à annoter un corpus de messages collectionnés, Reynolds, Kontostathis & Edwards (2011) et Sood, Antin & Churchill (2012b) ont essayé une méthode qui s'appelle 'externalisation ouverte' (*crowdsourcing*). Cette pratique fait appel aux connaissances des internautes intéressés pour résoudre un problème particulier, pour donner un avis, … Le site internet *Amazon's Mechanical Turk*[15], par exemple, permet aux chercheurs de formuler une tâche : dans ce cas-ci l'étiquetage des entrées et la reconnaissance de cas de cyberharcèlement parmi les entrées d'un corpus. Les utilisateurs d'*Amazon's Mechanical Turk* reçoivent une rémunération ($0,05) pour chaque tâche qu'ils effectuent correctement. Pour vérifier les compétences des utilisateurs, les chercheurs ont formulé des questions auxquelles ils ont données une réponse eux-mêmes. Au cas où un utilisateur donnerait deux fois une réponse fautive, toutes ses données étaient écartées du corpus. Ce mode de travail a plusieurs avantages : premièrement, il y a plusieurs annotateurs, ce qui augmente l'exactitude des étiquettes. Vu que les gens ont également du mal à décider de la nature offensive d'un message, les différentes perspectives garantissent un bon échantillonnage de la population. Deuxièmement, on garde seulement les messages du corpus qui ont été considérés comme exemples typiques de cyberharcèlement par deux tiers des annotateurs. Et dernièrement, un grand nombre de messages sont annotés dans un cadre de temps restreint.

Dans ce qui suit, nous analyserons un exemple d'un cas de cyberharcèlement pris par hasard dans *YouTube*. Nous déterminerons les traits pertinents et communs à chaque cas de cyberharcèlement[16]. La vidéo choisie dans *YouTube* est un enregistrement d'un discours du président Barack Obama lors des campagnes électorales américaines en 2012.

U_1 : Obama seems to be on the right track with what he wants to do to fix the economy. He just cant [*sic*] seem to get any help at all from Congress. I know the majority Republicans in Congress set out to see that Obama would fail before the inauguration even took place. They did'nt [*sic*] try to work with this President mainly I think was because of his race. I think Obama is at least for the average man or woman all Romney is for is the rich folks like himself.

U_2 : Fix the economy? He is presiding over the worst growth since the Great Depression. You are just kidding yourself if you think this Statist will fix the economy. He wants a diminished America. Massive debt + Slow growth + High spending kills empires. I expect your age exceeds your IQ.

U_3: My god you are stupid.

U_2: My comment is accurate and True.

U_3: No, your dumbass opinion is not accurate, not true, thus not making it a fact.[17]

[15] Amazon's Mechanical Turk. https://www.mturk.com/mturk/welcome [1/10/2012].

[16] Nous baserons notre analyse sur toutes les études déjà publiées sur la détection automatique de cas de cyberharcèlement. Nous avons utilisé surtout le corpus de Bayzick, Kontostathis & Edwards (2011) pour vérifier nos assertions : MySpace Group Data Labeled for Cyberbullying. http://www.chatcoder.com/DataDownload [1/10/2012].

[17] ABCNews, *President Barack Obama DNC Speech Complete: Romney in 'Cold War Mind-Warp' - DNC 2012*. [consultable en ligne sur :] http://www.youtube.com/watch?v=Hd8MFmUDbg4 [26/11/2012].

La première observation que nous nous sommes faite, c'est que nous voyons clairement que les divers aspects de la description du cyberharcèlement sont présents dans cette conversation : le harcèlement est répétitif dans le temps, le cyberharceleur a l'intention de blesser sa victime et – bien que ce ne soit pas un facteur déterminant - aucun utilisateur ne publie sa vraie identité.

Entrons dorénavant dans les détails de la conversation. La première réponse (de l'utilisateur U_2) contient la phrase « *I expect your age exceeds your IQ* » ('Je suppose que ton âge dépasse ton QI'). Dans ce cas-ci, il ne s'agit évidemment pas d'un compliment. L'utilisateur U_2 emploie une expression sarcastique (et sans injures) pour blesser sa victime (U_1). À cause des expressions sarcastiques, beaucoup de gens ont des difficultés à déterminer si un message est vraiment offensif ou menaçant. Parfois un énoncé est ambigu et peut donner lieu à plusieurs interprétations: une conversation entre bons amis qui se taquinent pourrait ressembler à un cas de cyberharcèlement, vu qu'ils s'insultent, mais ils le font de façon affective (voir *infra* 1.1). Il est également possible que quelqu'un renforce son opinion en se servant d'injures (Yin e.a., 2009 ; Dinakar e.a., 2012), sans s'adresser directement à autrui. C'est la raison pour laquelle Yin e.a. (2009) subdivisent les messages de cyberharcèlement en trois catégories : (1) les messages offensifs qui visent une personne, (2) les messages sarcastiques (dits aussi 'poliment' offensifs) qui visent une personne, et (3) les messages négatifs qui ne blessent personne en particulier (par exemple quelqu'un qui est mécontent d'une entreprise et qui véhicule ses émotions). Nous nous intéresserons dans notre mémoire principalement aux deux premières situations.

Après ces considérations, reprenons l'analyse de la conversation : la contre-réponse de l'utilisateur U_3 illustre un autre aspect, à savoir l'emploi répétitif du pronom personnel « *you* » ('tu'). Étant donné que le cyberharceleur adresse la parole directement à sa victime et qu'il veut la blesser, il utilise la forme de la deuxième personne singulier pour souligner l'individualité de l'injure (« c'est uniquement toi qui as tel défaut ») (Dinakar e.a., 2012). En plus, un cyberharceleur insulte souvent sa victime en se moquant de sujets qui font partie du caractère ou de l'apparence de la victime ; dans notre exemple, l'utilisateur U_3 se moque de l'intelligence de l'utilisateur U_2 (Dinakar e.a., 2012).

Soulignons aussi qu'à partir de cette entrée le thème du premier message n'est plus respecté : si le premier message commente encore le contenu de la vidéo, les commentaires qui suivent n'ont plus rien à voir avec les propos du président des États-Unis. Yin e.a. (2009) soutiennent que les cas de cyberharcèlement rompent souvent brusquement la conversation qui se déroule.

Le quatrième énoncé de notre exemple montre un fait que nous n'avons rencontré que dans une seule étude. Ce que Bayzick, Kontostathis & Edwards (2011) incluent parmi les caractéristiques du langage de cyberharcèlement, c'est l'emploi (exagéré) de majuscules. Souvent, les majuscules indiquent l'élévation de la voix, comme c'est aussi le cas dans les œuvres romanesques. L'emploi implique que l'utilisateur est étonné ou en colère, deux émotions qu'on peut facilement mettre en relation avec le cyberharcèlement. Ici, il ne s'agit que d'un seul mot écrit avec une majuscule, là où l'on ne s'y attend pas (« *True* », 'Vrai'). Peut-être que c'est simplement une faute de frappe ou que l'auteur voulait insister sur la vérité de son énoncé, sans que ce soit vraiment un trait pertinent dans notre exemple.

Enfin, le dernier commentaire de l'extrait confirme un fait soulevé par plusieurs chercheurs (Yin e.a., 2009 ; Dinakar e.a. 2012 ; Sood, Antin & Churchill, 2012a): la présence d'injures (« *dumbass* », 'idiot') est un trait

pertinent pour détecter des cas de cyberharcèlement. La détection d'injures est donc d'une grande importance, parce qu'elles sont de bons indices de cyberharcèlement. Il ne suffit toutefois pas de chercher uniquement les injures, puisque nous avons vu qu'un cyberharceleur peut utiliser aussi le sarcasme pour médire de quelqu'un.

Toutes ces observations montrent qu'il est possible d'automatiser la détection de cas de cyberharcèlement. Il y a assez de traits linguistiques que les techniques informatiques peuvent analyser. Tout de même il reste plusieurs difficultés à surmonter. En cherchant un exemple typique d'un cas de cyberharcèlement, nous avons rencontré deux nouveaux problèmes: 1) les messages suivent l'ordre chronologique et non pas l'ordre logique, et 2) comment sait-on si le cyberharceleur s'adresse à une certaine personne ou à une entité générale (telle qu'une entreprise, l'école, …)?

Abordons d'abord le problème de la dispersion des commentaires, ce qui est surtout remarquable dans *YouTube*. Les messages postés dans ce site internet suivent l'axe temporel. Par conséquent, les différentes entrées qui appartiennent à un même cas de cyberharcèlement sont dispersées dans les pages des commentaires (c'est-à-dire, d'autres commentaires interrompent la conversation). Heureusement, *YouTube* permet de relier les divers commentaires d'une même conversation, sous forme des soi-disant fils (*threads*).

Les autres réseaux sociaux populaires respectent davantage l'ordre logique : soit les commentaires sur un thème spécifique sont regroupés tout de suite dans un fil (c'est le cas pour *Facebook, Formspring.me, MySpace*), soit les auteurs des commentaires renvoient dans leurs messages aux personnes concernées (entre autres dans *Twitter*, avec l'abréviation RT : « *reply to/ retweet* », 'réponds à' ou l'arobase).

Cette dernière réflexion nous amène à l'autre difficulté à résoudre (qui n'est cependant pas très claire dans notre exemple) : le cyberharceleur s'adresse-t-il à une victime directe ou à une troisième partie (Sood, Churchill & Antin, 2012b)? La reconnaissance des victimes sert surtout à mieux intervenir lors de la détection d'un cas de cyberharcèlement. La même chose vaut d'ailleurs pour les cyberharceleurs: afin de les faire taire ou de leur expliquer que leur conduite est inacceptable, le système doit être capable de reconnaître les expéditeurs et les destinataires d'un message.

Pour les gens il semble que ces deux problèmes soient banals, mais ils sont d'une grande importance pour un système informatique. Surtout parce que ce n'est pas le but de simplement supprimer les messages de cyberharcèlement ; nous voulons faire réfléchir les cyberharceleurs sur leurs actes pour éviter qu'ils répètent leur comportement transgressif dans l'avenir.

Le système doit donc d'abord détecter des cas de cyberharcèlement et puis décider s'il a affaire avec un message d'un cyberharceleur ou d'une victime. Pour évaluer les performances du système, la linguistique informatique fait usage de trois notions importantes. Nous les présenterons dans la partie suivante.

2.1.1 Les performances du système : la précision, le rappel et le F-score

Avant de passer aux différentes techniques informatiques pour la détection de messages de cyberharcèlement, arrêtons-nous ici un instant sur les notions importantes de rappel, de précision et de F-score.

Le rappel représente le pourcentage des messages retenus correctement comme des cas de cyberharcèlement (cas positifs) à l'égard de tous les messages positifs dans un corpus.

La précision, par contre, donne le rapport des messages classés correctement au regard de tous les documents considérés comme des cas de cyberharcèlement.

On emploie le rappel et la précision pour calculer la moyenne harmonique des deux valeurs : le F-score (ou la F-mesure) égale le double du produit du rappel et de la précision, divisé par la somme des deux valeurs.

$$(1) \quad F\text{-}score = \frac{2 \cdot (précision \cdot rappel)}{précision + rappel}$$

Toutes les techniques de la linguistique informatique présentées dans les parties suivantes visent à créer un système semi-automatique par l'apprentissage supervisé : à partir d'un corpus annoté, les chercheurs apprennent au logiciel les traits pertinents pour développer ainsi un logiciel qui puisse assister un modérateur.

Le choix du type de système (automatique ou semi-automatique) a un effet considérable sur les performances du système. Sood, Churchill & Antin (2012b) énumèrent les répercussions pour le rappel et la précision selon le système choisi. Pour un système semi-automatique, on préfère le rappel à la précision: il vaut mieux signaler tous les cas retenus positifs (même s'ils sont douteux), puisque c'est au modérateur de prendre une décision sur la gravité du contenu. Même en ne sélectionnant que les cas jugés dangereux, la tâche du modérateur reste insurmontable. Une piste de recherche à suivre est une classification selon la gravité des insultes (Reynolds, Kontostathis & Edwards, 2012, voir aussi *infra* 2.2.1).

Si l'on veut concevoir un système tout à fait automatique, on doit préférer la précision au rappel. C'est que le système ne peut pas supprimer les commentaires qui ne sont pas vraiment offensifs, et il élimine donc seulement les plus graves cas de cyberharcèlement. Vu qu'il n'y a plus de modérateur qui prend la décision finale, il n'est pas conseillé que le système écarte automatiquement les cas douteux.

Outre qu'évaluer les performances du système, il faut voir quel modèle d'apprentissage supervisé donne les meilleurs résultats. Presque toutes les études estiment que la classification binaire du texte fonctionne le mieux sur la base de machines à vecteurs de support (SVM). Une machine à vecteurs de support crée un espace à plusieurs dimensions dans lequel elle regroupe les éléments similaires. Les divers axes représentent différents traits (injures +/-, pronoms personnels +/-,…), selon lesquels sont rassemblées les entrées d'un corpus. Après, une formule mathématique calcule la marge maximale entre la frontière et les données, de sorte que les cas positifs et négatifs soient bien séparés et que le système puisse apprendre quels sont précisément les traits qui permettent de détecter un cas de cyberharcèlement. L'algorithme décide s'il a affaire avec un cas de cyberharcèlement (positif, +) ou non (négatif, -). Selon trois recherches, la distribution binaire (+/-) serait plus efficace par rapport à l'emploi de plusieurs classes (Dinakar e.a.,2012 ; Sood, Antin & Churchill, 2012b; Xu e.a., 2012a). Nous aussi envisagerons un système qui enchaîne plusieurs classifications binaires.
Contrairement à toutes les autres études, Reynolds, Kontostathis & Edwards (2011) ont trouvé que la machine à vecteurs de support a le résultat le plus faible. Selon elles, un arbre de décision est plus efficace et solide (après huit étapes de vérification). Un arbre de décision crée un schéma à suivre pour classer un message donné. On peut le comparer avec une clé de détermination, qui donne des traits discriminants pour distinguer

entre plusieurs possibilités. Reste maintenant à faire plus de recherches pour voir lequel des deux modèles, SVM ou un arbre de décision, donne les meilleurs résultats.

Dans les parties suivantes nous donnerons un aperçu des techniques employées par les chercheurs pour construire un système semi-automatique qui détecte les cas de cyberharcèlement. Nous ferons un parcours du niveau morphosyntaxique au niveau métatextuel, en précisant et en évaluant toujours les performances des techniques proposées.

2.2 Les techniques informatiques au niveau morphosyntaxique

L'emploi (exagéré) d'injures a paru un trait très important dans notre analyse d'un cas de cyberharcèlement. Nous examinerons dans cette partie la détection d'injures, qui pose plus de problèmes qu'on ne soupçonne en première vue. Les injures ne servent pas seulement à détecter le cyberharcèlement ; elles donnent aussi une idée de la gravité du cyberharcèlement. Nous proposerons donc une méthode de classification à l'aide des injures. En plus, nous introduirons une autre approche plus large, à savoir la détection de mots fréquents (pas forcément injurieux) qui indiquent du cyberharcèlement et comment un système informatique arrive à trouver ces mots clés.

2.2.1 Une liste d'injures

Étant donné que beaucoup de messages de cyberharcèlement contiennent des injures[18], il est intéressant d'intégrer un algorithme qui les détecte. Pour y parvenir, il faut d'abord énumérer les injures les plus fréquentes dans des messages postés en ligne. Il y a plusieurs sites internet qui ont essayé de créer un inventaire d'injures[19]. Néanmoins, Sood, Antin & Churchill (2012a) posent qu'une simple énumération n'est pas efficace. Après un examen de quelques systèmes standard qui se basent sur des listes d'injures prises directement dans un site internet, elles ont conclu que ces systèmes ne trouvent que la moitié des injures présentes dans un corpus : la précision est de 52,8 % et le rappel de 40,2% (Sood, Antin & Churchill, 2012a). Nous expliquerons dans ce qui suit pourquoi les deux pourcentages sont tellement décevants.

Premièrement, l'orthographe peut varier : la rapidité avec laquelle un utilisateur dactylographie fait qu'il fait parfois des fautes de frappe. Elles peuvent être ou bien accidentelles (*siht, dumass*) ou bien destinées à déguiser intentionnellement l'injure (*@sshole, m*therf*cker, f###, $h!t*). Les procédés normaux ne réussissent pas à trouver ces injures altérées, d'où vient le pourcentage peu satisfaisant de la précision. Un bon moyen pour résoudre ces modifications est l'algorithme qui porte le nom 'la distance de Levenshtein'. Cet algorithme compare une chaîne standard (une phrase entière ou un mot unique) avec une chaîne modifiée et calcule combien de suppressions, d'ajouts ou de remplacements il y a eu dans la chaîne modifiée (Ptaszynski e.a., 2010). La valeur totale que donne la distance de Levenshtein est la somme de tous les changements. Sood, Antin & Churchill (2012a) soutiennent que le nombre de marques de ponctuation dans une chaîne donnée égale le nombre de suppressions, insertions ou déplacements. En mesurant la similarité des deux chaînes, il n'est pas nécessaire d'inventorier toutes les différentes variantes orthographiques possibles. La difficulté reste

[18] La présence des injures est plus significative que la fréquence (Sood, Antin & Churchill, 2012b).
[19] NoSwearing. http://www.noswearing.com [25/10/2012].
Dictionnaire des jurons. http://www.defoulodrome.com/jurons [25/10/2012].

de distinguer entre les divers emplois des signes de ponctuation. Par exemple, l'arobase peut indiquer soit une réponse, soit un renvoi à un autre utilisateur : @john vs @ss; de même le croisillon a différentes significations :

(1) *#tvv* (un mot-dièse qui regroupe les messages qui portent sur *The Voice Vlaanderen*) ;
(2) *grmbl#!@&* (renforce un énoncé) ;
(3) *f#ck* (injure).

L'intégration de la distance de Levenshtein augmente considérablement la précision; le rappel par contre dépend de l'exhaustivité de la liste d'injures de base (Sood, Antin & Churchill, 2012a).

Deuxièmement, les argots des jeunes changent énormément vite, ce qui crée un va-et-vient d'expressions en vogue. Les listes d'injures doivent donc régulièrement être mises à jour, pour y ajouter les nouvelles tournures qui se sont introduites dans le langage des jeunes (Sood, Antin & Churchill, 2012a).

Troisièmement, certains mots blessent seulement quand ils se trouvent dans des contextes spécifiques. Prenons par exemple l'expression *sale chienne* : ce groupe de mots a une autre nuance quand il paraît dans un forum dédié à des maîtres de chiens, par rapport à quand il sert de commentaire sur une photo d'une jeune fille. Puisque chaque réseau social a d'autres conditions d'utilisation, il serait utile d'établir d'abord une liste générale que les modérateurs peuvent adopter et adapter selon leurs besoins (Sood, Antin & Churchill, 2012a). Ainsi, pour une de leurs expériences, Dinakar e.a. (2012) ont extrait les termes pertinents dans des classes prédéterminées (l'intelligence, l'orientation sexuelle et la race ; voir aussi *infra* 2.3.1) pour rédiger des listes spécifiques. En comparant les résultats de la détection à partir d'une liste générale avec ceux des listes spécifiques, les chercheurs ont constaté que le classement catégoriel donne en général un meilleur F-score, avec un maximum de 0,77 (contrairement au 0,63 d'une classification générale). Il semble que la détection du cyberharcèlement soit un problème lié au domaine, c'est-à-dire la détection améliore si on connaît le thème d'un message et qu'il existe une liste spécifique pour ce thème. Ceci vaut surtout pour les sites internet qui offrent la possibilité de commenter un article, une vidéo, une photo…

Au lieu d'uniquement détecter les injures, Yin e.a. (2009) conseillent de prendre en considération l'emploi des pronoms personnels (surtout ceux de la 2^{ième} personne singulier). Xu e.a. (2012b) ont suivi cette idée et ont abouti à un F-score de 0,77 en faisant un mélange de monogrammes (un mot) et de digrammes fréquents (une combinaison de deux mots telle que *you […],[…] yourself,* etc.). Ptaszynski e.a. (2010) sont allés encore plus loin en ajoutant à chaque injure les informations grammaticales. Leur dictionnaire (y-compris l'étiquetage grammatical et la distance de Levenshtein) a donné un F-score de 0,843. Il faut cependant noter qu'ils ont construit leur dictionnaire à partir des injures les plus fréquentes dans leur corpus (et non pas en partant d'une liste générale trouvée sur Internet).

Outre qu'indiquer la présence de cyberharcèlement, les injures peuvent aussi aider à faire un classement des cas les plus urgents. Reynolds, Kontostathis & Edwards (2011) ont lié à chaque injure un 'niveau de criticité'. Cette méthode a l'avantage de classer les messages par degré d'urgence, de façon que les modérateurs ont l'occasion d'intervenir à temps dans les situations les plus critiques (Ptaszynski e.a., 2010). Reynolds, Kontostathis & Edwards (2011) sont parties d'une liste d'injures et de jurons. Dans leur étude, les traits étaient

les suivants : le nombre absolu et relatif de mots négatifs qui paraissent dans un message, un niveau de criticité (mot neutre – mot blessant – mot extrêmement blessant ; exprimé en nombres de 100 à 500) et une indication de la présence ou de l'absence du cyberharcèlement. Un algorithme calcule alors le degré d'urgence d'un message *i* en additionnant le nombre *k* des injures *n*, multipliées par leurs niveaux de criticité respectifs (de 100 à 500), et divisé par le nombre total *k* des mots dans ce même message *i*. Plus le résultat est élevé, plus l'énoncé est blessant.

$$(2)\quad degré\ d'urgence_i = \frac{\sum_k (n_i \cdot niveau\ de\ criticité)}{\sum_k n_{k,i}}$$

La détection d'injures paraît un bon point de départ, mais soulignons que 42 % des données dans l'étude de Sood, Antin & Churchill (2012a) contiennent des injures qui ne visent personne en particulier. Une discussion animée ou une opinion passionnée sont parfois accompagnées d'injures en guise de renforcements. La moitié des énoncés détectés ne seraient donc pas de cas de cyberharcèlement (c'est ce qu'on appelle en linguistique informatique 'les faux positifs'). Il faut par conséquent intégrer d'autres moyens pour améliorer la performance du système. Au niveau morphosyntaxique, on peut par exemple rédiger une liste de mots clés, qui inclut des injures ainsi que des termes fréquents (non-injurieux) dans des messages de cyberharcèlement.

2.2.2 Une liste de mots clés

Comment un système informatique peut-il déterminer quels sont les mots clés qui indiquent le cyberharcèlement ? La recherche d'information utilise souvent une méthode statistique appelée TF-IDF. Le score TF-IDF (*term frequency – inverse document frequency*) s'avère utile pour élaborer une liste de mots clés – qui contient des injures ainsi que des mots non-injurieux (Sood, Churchill & Antin, 2012b ; Yin e.a., 2009) . Le formule du score TF-IDF est le suivant :

$$(3)\quad TF \cdot IDF_{i,j} = \frac{n_{i,j}}{\sum_k n_{kj}} \cdot \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

La fréquence du terme (TF) est le nombre d'occurrences *n* d'un terme *i* dans un document *j* divisé par le nombre *n* total de tous les mots *k* dans le document *j*. La fréquence inverse de document (IDF) calcule le logarithme de l'inverse de tous les documents *j* dans lesquels paraît le terme *i* divisé par le nombre total *D* de tous les documents qui font partie du corpus.

Concrètement, supposons que le terme *salope* paraisse deux fois dans un document de cent mots. Ce document appartient à un corpus d'un million de documents, dont cent contiennent effectivement le mot *salope*. La fréquence du terme égale 0,02 (2/100). La fréquence inverse de document égale 4 (= $log(10^6/10^2)$). Le score TF-IDF fait donc 0,08 (= 0,02 · 4). Plus le terme est un bon discriminateur, plus le score TF-IDF est élevé. Par exemple, un mot qui paraît également deux fois dans un texte de cent mots, mais dans dix mille documents qui font partie d'un corpus d'un million de documents, a un score de seulement 0,04.

En calculant tous les scores TF-IDF d'un corpus constitué uniquement de messages de cyberharcèlement, le logiciel peut s'entraîner à ne trouver que les mots qui ont un score TF-IDF élevé, vu que ceux-ci sont de bons

discriminateurs pour le cyberharcèlement. Yin e.a. (2009) remarquent que le TF-IDF donne de meilleurs résultats par rapport aux n-grammes et aux injures isolées, mais tout comme la détection des injures, il faut combiner une liste de mots clés avec d'autres techniques informatiques.

2.3 Les techniques informatiques au niveau sémantique

Comme nous venons de constater, une liste d'injures et une liste de mots clés offrent une bonne base, mais elles ne parviennent pas à détecter tous les cas de cyberharcèlement. Pour augmenter le rappel ainsi que la précision du système, nous présenterons ici deux méthodes de la linguistique informatique qui utilisent des données sémantiques, à savoir la détection du thème et la reconnaissance d'entités nommées.

2.3.1 La détection du thème

La détection du thème (*topic detection*) n'a pas été étudiée intensivement dans le cadre de la détection automatique de cas de cyberharcèlement, bien qu'elle puisse être un bon outil. Étant donné que le cyberharceleur veut blesser sa victime (le plus sérieusement possible), il l'attaque sur des points qui font partie de la personnalité ou de l'apparence de la victime. Dinakar e.a. (2012) signalent qu'il y a quatre grands thèmes récurrents avec le cyberharcèlement: 1) l'apparence physique, 2) la race et la culture, 3) l'orientation sexuelle et 4) l'intelligence. La majorité des insultes s'alignent sur l'un de ces thèmes; si le thème d'un message correspond donc à un de ces quatre, on a de grandes chances d'avoir affaire avec un cas de cyberharcèlement.

Xu e.a. (2012b) ont fait d'autres distinctions, optant pour des qualifications plus générales: sentiments, suicide, famille, école, harcèlement verbal et harcèlement physique. Contrairement à la recherche de Dinakar e.a. (2012), ils n'ont pas uniquement travaillé sur des messages des cyberharceleurs ou des victimes. Ils ont simplement cherché les messages qui contenaient les mots « *bully*, *bullied* ou *bullying* ». En conséquence, ils ont également pris en considération les messages des reporters et des témoins, d'où cette autre classification. C'est que les témoins et les reporters racontent un cas de harcèlement classique à leurs amis en ligne. Cela explique la présence des thèmes « école », « harcèlement physique » et « harcèlement verbal ».

Jusque maintenant les chercheurs n'ont pas prêté beaucoup d'attention à la détection du thème dans des messages singuliers. Ceci ne signifie pas qu'ils n'emploient pas cette technique ; ils l'utilisent plutôt pour déterminer la similarité du thème d'un ensemble de messages sur le plan discursif (voir *infra* 2.4.2).

2.3.2 La reconnaissance d'entités nommées

Pour une intervention appropriée et personnelle dans un cas de cyberharcèlement, il est important de savoir quels sont exactement les rôles des personnes engagées dans une conversation en ligne. Remarquons cependant que ces rôles peuvent changer d'une conversation à une autre – ou même dans un seul discours. Il vaut donc mieux étudier plusieurs messages qui se succèdent afin d'aboutir à une attribution plus correcte des rôles (Xu e.a., 2012b). Les médias sociaux et les forums connaissent une organisation qui suit l'axe temporel : les commentaires des internautes ne sont pas rangés de façon logique, mais de manière temporelle. Ceci fait qu'une discussion est parfois dispersée dans les différentes pages d'un site internet. Pour reconstruire tout un cas de cyberharcèlement, on peut essayer de retrouver les messages successifs à l'aide des entités nommées[20].

[20] Une entité nommée est un substantif qui désigne une personne, un nom de lieu, un nom de personnage, un nom d'une entreprise ou quelque autre nom propre que ce soit. En général, on les écrit avec majuscule.

Il n'est pas surprenant que cette tâche ne soit pas facile, surtout parce que les corpus sont pleins d'entités nommées et que les algorithmes ne fonctionnent pas bien pour des textes courts (Ritter e.a., 2011).

Xu e.a. (2012b) ont distingué cinq classes qui s'accordent avec les rôles lors d'un cas de cyberharcèlement, à savoir l'accusateur, le cyberharceleur, le reporter, la victime et autre. D'abord, ils attribué un rôle à l'auteur d'un message. Il apparaît que l'accusateur et le cyberharceleur sont les plus difficiles à reconnaître : ils sont fréquemment confondus avec le rôle du reporter. Aussi, concluent les chercheurs, est-il possible qu'il y ait certains traits linguistiques communs à l'accusateur, le cyberharceleur et le reporter. Puis, ils ont essayé de classer dans une des cinq classes toutes les autres entités nommées et les pronoms personnels qui se trouvent au sein d'un message. En employant les valences des verbes, ils ont pu reconstruire les divers objets grammaticaux d'un syntagme verbal. Dans l'exemple *Béatrice a dit à Dante qu'il est un idiot*, Béatrice (sujet) est la cyberharceleuse, Dante (COI) la victime et l'auteur du message le reporter. La difficulté que Xu e.a. (2012b) ont dû surmonter était de séparer les entités générales (*les gens, la police, Dieu,…*) de noms de personnes. Ils ne pouvaient pas simplement supprimer tous les substantifs, vu que beaucoup de noms communs renvoient à des individus. De fait, ils ont rencontré des difficultés avec la classification de substantifs qui désignent une personne (tels que *mon frère, le baby-sitter, son professeur …*). Par conséquent, l'attribution des rôles n'est pas sans erreurs : bon nombre de victimes et de cyberharceleurs ont été mal classés et considérés comme 'autres'.

Ptaszynski e.a. (2010) ont eux aussi entraîné leur système dans la recherche de noms propres, à partir d'informations personnelles (telles que l'adresse, le numéro de téléphone, les indices de l'identité,…). Ils ne sont pas allés jusqu'à attribuer des rôles à ces noms. Pour eux, les informations supplémentaires ne servent qu'à trouver des cas de cyberharcèlement : souvent un cyberharceleur publie des (coor)données privées ou des informations sensibles sur sa victime.

Sood, Churchill & Antin (2012b) ont seulement vérifié si l'auteur d'un certain message renvoie à un auteur d'un message précédent, sans que le rôle de l'auteur n'ait d'importance. La présence de digrammes racinées[21] a fourni les meilleurs résultats : un F-score de 88,21 %. Ceci montre que la reconstruction d'une conversation est vraiment possible à partir des noms des utilisateurs.

La reconnaissance d'entités nommées sert surtout à mieux estimer la portée d'un cas de cyberharcèlement. Si un modérateur peut juger sur tous les messages d'une conversation qui se déroule entre les mêmes utilisateurs, il est moins enclin à se tromper (souvenons-nous des amis qui se taquinent). L'attribution des rôles garantit aussi de meilleures possibilités d'intervention : le modérateur sait tout de suite qui a besoin d'aide, et qui, en revanche doit, être banni du réseau social. Puisque le modérateur a besoin d'un contexte, le système devra aussi tenir compte des messages précédents et suivants. Les techniques aux niveaux discursif et pragmatique méritent donc plus d'attention.

2.4 Les techniques informatiques aux niveaux pragmatique et discursif

Il est clair qu'un cas de cyberharcèlement ne contient pas toujours les mêmes mots ou structures. Nous avons vu, par exemple, que la présence ou l'absence d'injures ne dit rien sur le caractère blessant d'un message. Dans une conversation réelle, la façon dont on dit quelque chose détermine la gravité d'un insulte. Pensons à deux

[21] Les digrammes racinés sont, dans ce cas, deux mots qu'on met à leur forme de base. Par exemple : '*harcèle-moi*' = 'harceler' + 'moi'

amis qui s'insultent, mais ils le font en riant de sorte que ils savent qu'ils plaisantent. Les véritables messages de cyberharcèlement ont une grande caractéristique en commun : ils veulent jouer sur les émotions de la victime. Cette observation sera notre point de départ pour cette partie.

Les techniques informatiques au niveau pragmatique comprennent l'analyse sentimentale et la détection du sarcasme. Ces deux tâches informatiques visent à 'calculer' la valence (ou la polarité) d'un énoncé : s'agit-il d'un message positif, négatif ou sarcastique ? Nous partirons de l'idée que les énoncés soit négatifs, soit sarcastiques et négatifs à la fois pourraient indiquer un cas de cyberharcèlement.

Au niveau discursif, les techniques informatiques analysent plusieurs messages en même temps. Comme l'a montré notre analyse d'un cas de cyberharcèlement, un thème divergent est un bon indice pour le cyberharcèlement : un grand nombre de commentaires offensifs ne respectent pas le thème d'un message initial. Il est donc important de voir si les thèmes des commentaires s'accordent avec celui du premier message posté. Nous présenterons deux approches possibles pour effectuer cette comparaison.

Regardons d'abord comment la linguistique informatique est capable de 'lire' les émotions cachées dans un message grâce à une technique qui a reçu beaucoup d'attention ces dernières années.

2.4.1 L'analyse sentimentale

Nous avons souligné à plusieurs reprises qu'un cyberharceleur a l'intention de blesser sa victime. Pour cela, il utilise des phrases avec un contenu négatif. La victime réagit sur les messages de cyberharcèlement avec des phrases émotionnellement chargées (tristesse, colère, incompréhension…). Notre point de départ est qu'un message de cyberharcèlement se caractérise par un contenu négatif (soit de la part du cyberharceleur, soit de la part de la victime). Voilà pourquoi l'analyse sentimentale est intéressante : les algorithmes qui effectuent l'analyse sentimentale essayent d'attribuer un sentiment à un texte donné. Nous parlerons sous cette partie de deux diverses approches de l'analyse sentimentale : premièrement, l'analyse binaire qui détermine seulement la valence (positive ou négative) du contenu ; deuxièmement, l'analyse à plusieurs classes qui veut vraiment attribuer l'émotion correcte à un texte donné.

Une analyse sentimentale binaire veut classer les messages d'un corpus dans des classes prédéterminées, en l'occurrence les classes 'contenu négatif' et 'contenu positif'. Cette classification peut se faire de plusieurs façons. Bayzick, Konthostathis & Edwards (2011) ont développé un algorithme qui permet de détecter un contenu négatif dans un ensemble de messages collectionnés dans MySpace. Le logiciel s'appelle *BullyTracer* et repose sur une liste d'injures et de jurons fréquents. Le problème à résoudre est celui des nombreux faux positifs : le programme a des difficultés à classer correctement les messages positifs ou neutres, les classant trop souvent dans la classe 'contenu négatif' (*BullyTracer* a un calcul d'incertitude[22] de 58,63%).

Dinakar e.a. (2012) ont élaboré davantage l'emploi de l'analyse sentimentale : ils ont consulté *Ortony's Affective Lexicon,* un vocabulaire qui inventorie tous les mots désignant une émotion particulière. Les

[22] Le calcul d'incertitude est une notion en linguistique informatique qui sert à évaluer la performance d'un système . Pour calculer le calcul d'incertitude, il suffit de diviser les documents classés correctement par tous les documents dans le corpus. Dans notre mémoire, nous avons opté pour d'autres notions, pour la simple raison que le calcul d'incertitude ne dit pas grand-chose. Vu le nombre relativement limité de cas de cyberharcèlement dans les médias sociaux (ca. 5%), le système pourrait considérer tous les messages comme négatifs (il n'y a pas de cyberharcèlement) et aboutir à un calcul d'incertitude de 95 % (Sood, Antin & Churchill, 2012b).

chercheurs n'ont gardé que les termes liés à l'expression de la négativité : dès que le système détecte un des mots de la liste, il est probable que le message est négatif.

Remarquons que ces deux études utilisent en fait une méthode au niveau lexical pour faire des assertions au niveau pragmatique. Sood, Churchill & Antin (2012b) mentionnent que les listes de mots n'offrent jamais de détection solide. Elles critiquent principalement que la présence d'un certain mot dans un message n'implique pas forcément que le contenu soit négatif ou positif (prenons l'exemple des chercheuses : 'froide' dans une *boisson froide* ou une *femme froide*). Nous nous demandons s'il ne vaut pas la peine d'énumérer chaque fois deux mots (bigrammes) au lieu de mots simples (unigrammes), parce qu'ainsi certaines expressions gardent leur nuance émotionnelle (telle que *femme froide*).

Soulevons encore une autre réflexion. Nous avons déjà discuté une classification par degré d'urgence à l'aide des injures. Ne serait-il pas possible de classer les messages selon leur valence ? Plus la polarité d'un message tend à la négativité, plus il est urgent que le modérateur s'en occupe.

Certains chercheurs ne se contentent pas d'une classification binaire ; ils veulent savoir quelle émotion est précisément exprimée par un message pour garantir une meilleure détection de cas de cyberharcèlement.

Xu e.a. (2012a) ont constaté que les sept émotions les plus fréquentes dans leur corpus étaient: « colère, gêne, empathie, peur, fierté, soulagement et tristesse. » Comment faut-il classer un message dans une des sept catégories d'émotions? En premier lieu, les scientifiques ont cherché les mots pertinents qui sont associés aux émotions nommées ci-dessus à l'aide d'un dictionnaire de synonymes et la base de données lexicales *Wordnet*. À partir de leurs découvertes, ils ont rédigé une liste de termes fréquents. En deuxième lieu, ils ont effectué une recherche sur Internet pour trouver des documents en ligne contenant un des mots qui figuraient dans la liste préétablie. Les documents étaient surtout pris dans l'encyclopédie en ligne *Wikipedia* : les articles sur les émotions (par exemple « colère ») étaient les points de départ, les articles y liés ont servi à élargir le corpus.

Les chercheurs ont trouvé que 94 % des messages de cyberharcèlement dans leur corpus (pris dans *Twitter*) ne contenaient aucune trace des sept émotions prédéterminées. Cela veut dire que seulement 6 % de leurs énoncés montraient de l'émotion. Dans cet ensemble limité, ils ont vu deux grandes lignes de force : 1) l'empathie, la gêne et la fierté sont presque absentes dans le corpus ; et 2) certaines émotions sont liées à certains rôles fixes : les messages d'un accusateur montrent plus de peur et moins de colère ; ceux des victimes et des reporters sont plus tristes et soulagés.

Ptaszynski e.a. (2010) soutiennent aussi la thèse que les données blessantes contiennent moins d'émotions par rapport aux commentaires inoffensifs. Ils estiment que le cyberharceleur vise sa victime en lui disant des mots mûrement réfléchis, pour s'assurer que ses propos la blessent. Selon eux, un cyberharceleur est plus enclin à employer des émotions négatives (telles que la méprise ou la colère). À part la négativité, les chercheurs étaient étonnés de retrouver dans leur corpus aussi l'emploi d'affection. Cela peut s'expliquer par l'usage d'expressions sarcastiques, qui font semblant de créer une conversation entre amis.

Pour leur étude, Ptaszynski e.a. (2010) ont étudié un corpus de messages provenant d'un forum scolaire du Japon. D'abord, ils ont vérifié si les messages étaient émotionnellement chargés ou non. Pour aboutir à une classification automatique, ils ont utilisé un dictionnaire d'« émotèmes » et d'expressions sentimentales.

Par *émotème* ils ont compris « les interjections, les exclamations, les mots vulgaires et les expressions mimétiques » (ces dernières dans le cas du japonais, elles n'existent pas en français).

Puis, ils voulaient reconnaître de quel type d'émotion il s'agissait exactement. Pour cela, ils ont pris en considération les émotèmes, les émoticônes et le contexte dans lequel les expressions sentimentales paraissaient. À la lumière du contexte, les chercheurs ont souligné l'importance des changeurs de la valence du contexte (*contextual valence shifters* ou *CVS*), un phénomène que Polanyi & Zaenen (2006) avaient décrit. Les CVS sont de petits (ad)verbes qui changent la valence d'un énoncé (soit ils la renforcent, soit ils l'affaiblissent). La valence d'un énoncé est alors la somme des valences singulières : les émotèmes positifs ont une valeur de +1 ou +2 ; les émotèmes négatifs -1 ou (encore pire) -2.

Il y a trois grands types de CVS qui semblent intéressants pour l'analyse sentimentale de messages courts, à savoir les négations (*ne…personne, ne…jamais, ne…pas*,…), les modalisateurs intensifiants (*beaucoup, assez, un peu,…*) et certains verbes (*rater*, *manquer*, *ignorer*,...). Si un de ces types paraît dans une phrase, combiné d'une expression sentimentale, la valence de l'énonce doit être ajustée. C'est-à-dire, si dans un message quelqu'un met un terme négatif précédé d'un changeur de la valence du contexte, celui-ci rend le message (plus) positif (Polanyi & Zaenen, 2006). Par exemple :

(1) *Il n'est pas stupide.* (positif, puisque la négation nie la négativité)
(2) *Il est peu intelligent.* (négatif, parce que l'adverbe *peu* change la valence)
(3) *Il manque de l'intelligence.* (négatif, à cause du verbe *manquer*)

Cette observation nous offre un nouvel argument pour une liste composée de bigrammes. Vu qu'une négation invertit la valence d'une expression, ne vaut-il pas mieux de considérer les deux termes en même temps ? Prenons l'énoncé « *il est peu intelligent* » : *intelligent* donne normalement une valence positive à une phrase, *peu intelligent* en revanche donne l'idée contraire (et donc négative). De telle façon, le bigramme *peu intelligent* donne tout de suite la valence correcte de l'énoncé entier.

Il n'est pas étonnant que les CVS se rencontrent souvent dans les messages sarcastiques, qui visent à ridiculiser une personne en formulant le contraire de ce qu'on veut vraiment dire. Regardons donc plus en détail la détection du sarcasme dans les messages des médias sociaux.

2.4.2 La détection du sarcasme dans les médias sociaux

Nous avons vu plusieurs fois que le sarcasme négatif (ou insultant) se rencontre souvent dans des messages de cyberharcèlement. La notion de valence nous permet de mieux décrire ce que c'est que le sarcasme en termes de linguistique informatique. González-Ibáñez, Muresan & Wacholder (2011) définissent le sarcasme comme « un mécanisme qui transforme la valence d'un message positif ou négatif en son contraire », mais nous y ajouterions que cette transformation a lieu sans qu'on ait nécessairement recours à des injures. Il se peut que cela se produise par l'emploi des CVS, ou bien par une transformation des connaissances du monde. La question est maintenant de savoir comment un système informatique peut tenir compte des connaissances humaines et d'une culture générale partagées? Dinakar e.a. (2012) ont développé une méthode pour identifier les expressions qui s'opposent (ou pas) aux connaissances du monde. Avant d'entrer dans les détails de leur recherche, nous traiterons de la détection du sarcasme à l'aide de traits lexicaux et pragmatiques.

González-Ibáñez, Muresan & Wacholder (2011) ont essayé de détecter le sarcasme dans des messages de *Twitter*. Un des objectifs qu'ils s'étaient posés, était de séparer les messages non-sarcastiques (c'est-à-dire pleinement positifs ou négatifs) des messages sarcastiques. Ils ont constitué leur corpus en cherchant les mots-dièse *sarcasme*, les sentiments positifs (*joie, bonheur,…*) ou les émotions négatives (*colère, tristesse*,…). Ils ont choisi cette méthode de travail, parce qu'ils pensent que les auteurs des messages savent mieux que quiconque quand ils ont l'intention d'être sarcastiques.

Les chercheurs ont utilisé un dictionnaire de mots affectifs, une liste d'interjections, les émoticônes les plus populaires et les renvois à d'autres utilisateurs (par l'emploi de l'arobase : @[…]) comme traits pertinents. Selon leur expérience, les traits sélectionnés ne suffisent pas pour aboutir à une bonne détection du sarcasme : leur système ne réussit pas à classer correctement les messages sarcastiques dans la classe correspondante. Ce qui est intéressant, c'est que les annotateurs humains ont rencontré les mêmes difficultés que le système automatique. Ceci est dû au fait qu'il s'agit de messages uniques, sans contexte. La classification de messages contenant des émoticônes donnait le meilleur résultat, pour les hommes ainsi que pour le système. Il semble donc que les émoticônes fournissent des informations pragmatiques qui aident à mieux comprendre l'intention (et l'émotion) de l'auteur.

Pourquoi est-il tellement difficile de déterminer quand un message est sarcastique ? Nous avons déjà cité l'absence d'un contexte, mais il y a encore un autre aspect qui rend difficile la détection : les connaissances du monde. Un grand nombre d'utilisateurs (et locuteurs) font appel à leurs connaissances du monde lors d'une conversation : ils n'explicitent pas les informations qui leur paraissent évidentes (suivant une des maximes de la pragmatique, formulés par Grice). Si nous retournons à notre analyse d'un cas de cyberharcèlement dans la partie 2.1, nous voyons que le deuxième utilisateur a fait la remarque « *I expect your age exceeds your IQ* ». Grâce aux connaissances du monde, on sait que le quotient intellectuel moyen se situe généralement autour de 100. Tout le monde qui partage les connaissances de cet utilisateur comprend qu'il a l'intention de blesser l'autre, comme il existe peu d'internautes centenaires.

Dinakar e.a. (2012) sont les premiers chercheurs qui ont prêté attention à l'importance des connaissances du monde et à la façon dont on peut le simuler dans le domaine de l'intelligence artificielle. Dans leur étude, ils ont mis l'accent sur les insultes concernant l'orientation sexuelle. Ce type d'insultes intervertit souvent les stéréotypes liés aux deux sexes : les hommes homosexuels auraient des traits féminins et les femmes lesbiennes se conduiraient comme des hommes. Pour établir les rapports entre les sujets masculins et les 'attributs généralement masculins', les scientifiques ont utilisé *ConceptNet*[23], un logiciel qui décrit les relations qu'entretiennent les mots intégrés dans la base de données lexicales *WordNet*[24]. Il y a douze sortes de relations : *élément de, sert à, désire, appartient à, destiné à, est pour*… . Ainsi, on crée un réseau de mots au niveau de leur signification (un rouge à lèvres *est* un produit de cosmétique, *sert à* rougir les lèvres, *sert à* séduire,…), mais également au niveau de ses associations (en général, un rouge à lèvres *est* un produit *pour* les femmes). Dinakar e.a. (2012) ont complété les relations de *ConceptNet* en les enrichissant de données qu'ils

[23] ConceptNet. http://conceptnet5.media.mit.edu/ [21/01/2013].
Pour avoir une bonne introduction au logiciel, nous recommandons Liu & Singh (2004).
[24] WordNet. http://wordnet.princeton.edu/ [21/01/2013].

ont trouvées dans leur propre corpus de messages (de *Formspring.me*). Voici quelques exemples de leurs observations: une perruque *est pour* les filles, un toupet *est pour* les hommes,…

Leur expérience visait à vérifier si leur système donnerait des résultats comparables à ceux avec les opinions des annotateurs humains. Dans environ 70 % des cas, les messages qui renvoyaient respectivement à une fille ou un garçon étaient classés correctement. Après que le logiciel a déterminé 'le sexe de l'énoncé', il peut voir si celui-ci égale le sexe de la personne adressée. Sinon, on a probablement affaire à un cas de cyberharcèlement. Les 30 % des entrées qui étaient mal classées, recouvraient deux grands problèmes récurrents : en premier lieu, l'insuffisance de données (annotées) ou de réseaux conceptuels ralentit la création d'un meilleur système (voir *infra* 2.1 pour une discussion sur les corpus) ; en deuxième lieu, un message doit toujours être compris dans son contexte.

Jusqu'ici, nous avons traité les techniques informatiques qui détectent des messages uniques. Les conclusions des chercheurs montrent clairement qu'il faut également prendre en considération le contexte lors de la détection du sarcasme. Pour tenir compte des messages encadrés, nous proposerons deux différentes méthodes qui partent de l'idée de la similarité du thème.

2.4.3 La similarité du thème par rapport à un message initial et la contextualité

En discutant de la détection du thème, nous avons brièvement mentionné que cette technique prévalait surtout au niveau discursif. Bien que toutes les techniques, dont nous avons traitées, visent à trouver des messages uniques, le cyberharcèlement apparaît le plus souvent dans un contexte conversationnel. Notre exemple d'une conversation de cyberharcèlement montre que dans bien des cas un cyberharceleur interrompt brusquement une discussion qui respectait jusqu'à ce moment le thème d'un message initial. Surtout dans les forums, il vaut donc la peine de comparer le thème des commentaires avec celui du message au début.

Yin e.a. (2009) proposent de tenir compte de deux aspects contextuels, à savoir la notion de 'similarité du thème' ainsi que celle de la 'contextualité'. Par *similarité du thème* les chercheurs entendent que le même thème (et donc la même personne ou entité adressée) revient dans toute la conversation. Concrètement : quand dans une conversation quelqu'un parle du président des États-Unis, il est suspicieux qu'un autre utilisateur commence une phrase par un pronom personnel de la deuxième personne singulier.

La *contextualité*, par contre, ne part pas du thème du premier message, mais du principe que lorsqu'un utilisateur publie un message blessant, celui-ci est souvent suivi d'une accumulation de réactions négatives. La contextualité sert donc à déterminer la ressemblance des réponses entre elles. Un message blessant 'attire' des réactions qui ont également un contenu négatif. Yin e.a. (2009) ont pris pour leur étude un cadre de trois messages précédents et suivants. Leur système (qui comporte en outre une liste d'injures et de mots clés) a donné un F-score de 0,313.

Sood, Churchill & Antin (2012b : 280-281) envisagent une autre approche pour déterminer la similarité du thème. Elles sont parties de l'idée que la similarité de deux messages est mesurable, et par conséquent exprimable en pourcentages. Elles ont employé le score TF-IDF, une méthode statistique de pondération (voir *infra* 2.2.2), pour comparer les termes paraissant dans les commentaires ultérieurs avec ceux du message initial. À chaque mot, le système attribue une valeur selon la fréquence du mot dans le corpus. Si la somme de

toutes les valeurs des termes figurant dans un commentaire donné reste au-dessous d'un seuil prédéterminé, on pourrait avoir affaire à un cas douteux (spam ou cyberharcèlement). Plus la somme des scores TF-IDF est élevée, plus les thèmes des messages se ressemblent.

2.5 Les techniques informatiques au niveau métatextuel

Nous avons présenté les techniques informatiques qui peuvent effectuer les recherches nécessaires pour détecter les traits linguistiques du cyberharcèlement textuel dans les médias sociaux. Dans notre analyse d'un cas de cyberharcèlement (voir *infra* 2.1), nous n'avons pas tenu compte des traits métalinguistiques (tels que le sexe, l'âge, le lieu de résidence,…), simplement parce que les comptes des trois utilisateurs ne donnaient pas d'informations supplémentaires au moment de l'extraction dans *YouTube*. Cela ne veut cependant pas dire que les techniques computationnelles qui analysent les données métalinguistiques ne servent à rien lors de la détection de cas de cyberharcèlement. Peu d'études (en fait, nous n'en avons trouvé qu'une seule) se sont occupées de l'influence des données métatextuelles sur la détection de cas de cyberharcèlement. Dans ce qui suit, nous examinerons quel effet a le sexe d'un auteur d'un message de cyberharcèlement pour les performances d'un système (semi-)automatique.

2.5.1 Le sexe

En sociolinguistique les chercheurs ont publié des centaines d'articles qui portent sur l'écart linguistique entre les hommes et les femmes. Herring (2004) énumère les différences les plus remarquables entre le langage des deux sexes en ce qui concerne la communication en ligne. Les femmes trouvent important que tout le monde dans la communauté se sente à l'aise. Pour atteindre ce but, elles emploient des expressions plus affectives et appréciatives. Les hommes, en revanche, sont plus directs et n'hésitent pas à exprimer leur opinion en utilisant des termes forts (songeons à *flaming* : l'utilisateur poste des messages offensifs sans qu'il y ait une véritable raison). Vandebosch e.a . (2006 : 21) remarquent que les cyberharceleurs préfèrent cyberharceler leur victime directement ; les cyberharceleuses optent plutôt pour des méthodes indirectes. Selon les chercheurs, les filles auraient une attitude 'moins' violente, parce qu'elles sont (physiquement) moins fortes.

Comment ces caractéristiques langagières des sexes peuvent-elles aider la détection de cas de cyberharcèlement ? Dadvar e.a. (2011) ont inclus pour leur expérience les traits typiques au langage des hommes et des femmes d'un corpus collectionné dans *MySpace*. Après avoir séparé les messages écrits respectivement par les hommes et par les femmes, ils ont calculé la fréquence et la présence des insultes et des jurons pour chaque sexe. Il apparaît que la liste est significativement diverse. En intégrant ces données dans une machine à vecteurs de support (ensemble à l'emploi de pronoms et quelques mots clés), les chercheurs ont constaté une amélioration du F-score de 15 % par rapport à l'analyse sans distinction des sexes. Il faut quand même remarquer que le F-score général reste seulement de 0,23.

Cette étude montre qu'il vaut la peine de considérer également les données métalinguistiques. Vu que beaucoup de réseaux sociaux mettent en disponibilité un compte public qui contient des informations sur l'auteur d'un message, il est utile de classer d'avance un commentaire dans une classe restreinte (homme ou femme) pour adapter le mieux possible les techniques informatiques aux messages. Nous avons déjà insisté sur le fait que le cyberharcèlement est plus facilement détecte à l'aide de listes spécifiques. Ainsi, il serait par

exemple intéressant que le système tienne compte de la personnalité ou de l'âge : plus un cyberharceleur est âgé, plus grande est la chance qu'il emploie le sarcasme pour blesser sa victime. Soyons néanmoins prudents : les utilisateurs ne sont pas toujours honnêtes et souvent leur identité virtuelle ne correspond pas à la réalité.

2.6 Conclusions

Dans ce chapitre, nous avons essayé de fournir une réponse à deux questions de recherche : premièrement, nous nous sommes posé la question de savoir quels sont les traits linguistiques et/ou métalinguistiques pour automatiser la détection de cas de cyberharcèlement. Et deuxièmement, la suite logique : quelles sont exactement les techniques en linguistique informatique qui permettent de détecter les traits dégagés ?

Pour résoudre la première question, nous nous sommes occupés d'abord d'une analyse d'un cas de cyberharcèlement réel. Chemin faisant, nous avons dégagé plusieurs traits pertinents qui caractérisent un message de cyberharcèlement. Sur la base de ces observations, nous avons conclu qu'il est effectivement possible d'automatiser la détection de cas de cyberharcèlement.

À partir de cette conclusion, nous avons fait passer en revue plusieurs techniques computationnelles qui peuvent effectuer les différentes tâches de classification.

Au niveau morphosyntaxique, nous avons retenu comme importants les injures et les termes fréquents. À l'aide de la méthode statistique TF-IDF, il est possible de créer facilement une liste de mots clés. Nous avons également discuté de la rédaction d'une liste d'injures et la façon dont on peut augmenter la précision de la détection en employant l'algorithme de la distance de Levenshtein.

Au niveau sémantique, nous avons montré les possibilités (quoique peu utilisées par les chercheurs) qu'offrent la reconnaissance des entités nommées et la détection du thème. Les entités nommées servent surtout à mieux comprendre les rôles lors d'une conversation, c'est-à-dire qui est le cyberharceleur et qui la victime dans une conversation de cyberharcèlement. La détection du thème, par contre, paraît surtout valable au niveau discursif. Nous avons constaté que les messages de cyberharcèlement appartiennent généralement à quatre grands thèmes : l'apparence physique, la race et la culture, l'orientation sexuelle et l'intelligence.

Au niveau pragmatique, nous avons parlé de l'analyse sentimentale et de la détection du sarcasme. Les deux techniques tiennent compte de la valence d'un énoncé : elles veulent déterminer si un message est positif négatif ou sarcastique. Dans le cas de l'analyse sentimentale, nous avons soutenu la thèse que les cas de cyberharcèlement ont principalement un contenu négatif. Nous avons montré comment l'analyse sentimentale peut contribuer à une meilleure classification. Un problème en particulier a reçu plus d'attention, à savoir la détection du sarcasme. Outre que la négativité ou la positivité d'un message, nous avons introduit le côté sarcastique. À partir d'une base de données basée sur *ConceptNet*, un système informatique peut tenir compte des connaissances du monde. De telle façon, le système est capable de comparer les relations stéréotypées de concepts masculins et féminins (*le rouge à lèvres est pour une fille*) avec celles d'un message donné.

Au niveau discursif, nous avons insisté sur le fait qu'un message de cyberharcèlement fait généralement partie d'une conversation en ligne et aussi faut-il considérer également le contexte. Une approche en particulier nous a semblé intéressante : à l'aide du score TF-IDF, on peut mettre en rapport le thème d'un

commentaire avec celui d'un message initial. Si les deux n'ont pas de thèmes trop divergents, on a de grandes chances à avoir trouvé un cas de cyberharcèlement.

Au niveau métatextuel, nous avons mentionné les propriétés linguistiques de chaque sexe. Nous avons constaté que si l'on tient compte de ces différences en rédigeant une liste de mots clés, on aboutit à une détection plus nuancée.

Il était notre objectif de créer un système (semi-)automatique pour détecter des cas de cyberharcèlement textuel. Voilà pourquoi nous avons conçu ce petit schéma de différentes techniques informatiques qui permettent de construire un tel système automatique.

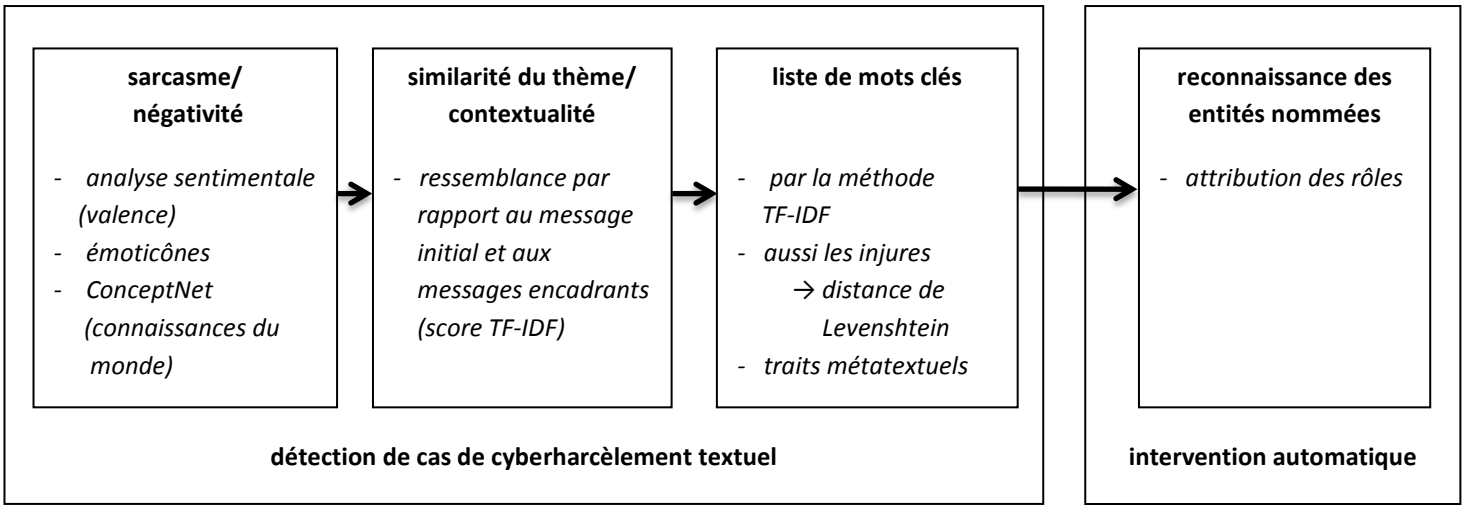

Figure 2: la détection (semi-)automatique de cas de cyberharcèlement textuel dans les médias sociaux

Regardons la figure 2 de plus près. Selon Sood, Antin & Churchill (2012b), les commentaires négatifs qui ne respectent pas le thème d'un message initial indiqueraient un cas de cyberharcèlement. C'est la raison pour laquelle nous avons mis la détection de la négativité et du sarcasme au premier rang, suivie du principe de la contextualité et de la similarité du thème. Nous avons élaboré la description des trois chercheuses en ajoutant la liste de mots clés (y-compris les injures et les traits métatextuels). Cette tâche est placée au troisième rang parce qu'elle permet au système de prendre une décision sur la criticité d'un message (voir *infra* 2.2.1), bien que nous croyions maintenant que la valence totale d'un message offre également un bon indice sur le degré d'urgence (plus le message est négatif, plus il est urgent que le modérateur en soit au courant). Ces trois tâches font toutes partie de la détection de cas de cyberharcèlement textuel.

Après la détection, c'est l'intervention qui suit. Un système automatique doit proposer les outils les plus adéquats pour aider le mieux possible les personnes concernées. Pour faire cela, il est indispensable que le système sache qui est le cyberharceleur et qui est la victime. Par conséquent, la reconnaissance des entités nommées est surtout valable pour l'intervention personnelle. La détection ainsi que l'intervention garantissent que le cyberharceleur et la victime se rendent compte de la portée du cyberharcèlement. Dans le troisième chapitre, nous poursuivrons cette piste et nous parlerons davantage des modes d'intervention automatique après la détection d'un cas de cyberharcèlement.

3. Modes d'intervention après la détection d'un cas de cyberharcèlement textuel

Une fois un commentaire d'un réseau social est signalé comme un cas de cyberharcèlement, le site internet le rapporte tout de suite au modérateur. Celui-ci doit décider s'il veut supprimer le message ou s'il le considère comme inoffensif. D'une manière ou d'une autre, l'auteur du message n'est pas mis au courant et, par conséquent, il ne se rend pas compte de son comportement transgressif. Il peut facilement continuer à écrire des commentaires blessants sans qu'il y ait des répercussions directes. Ne serait-il pas plus opportun d'intervenir directement et de montrer au cyberharceleur les conséquences de son message offensif, de façon que le cyberharcèlement s'arrête ? La même chose vaut d'ailleurs pour la victime : n'a-t-elle pas besoin d'aide directe après qu'elle a reçu un message blessant? Ce chapitre portera sur les différents modes d'intervention au moment où le système a détecté un cas de cyberharcèlement.

Dinakar e.a. (2012 : 13-18) donnent toute une liste de possibilités pour intervenir convenablement. Selon eux, l'intervention devrait entrer en action lors de l'action, c'est-à-dire le système doit immédiatement remédier à la conduite inappropriée d'un utilisateur. Ils proposent cinq modes d'intervention:

1) **Interruption**. L'interruption peut se faire de deux façons. Une première manière essayera de donner des signaux subtils pour que le cyberharceleur renonce à sa volonté d'envoyer le message. Dans un discours en personne, les interlocuteurs trahissent souvent leurs vraies émotions ou pensées à travers leur langage du corps. Ce sont des adaptations à peine consciemment remarquables : l'intonation change, le volume de la voix augmente, les yeux se rétrécissent,… Au lieu de ces indices corporels, le système de détection pourrait souligner les mots offensifs pour y attirer l'attention du cyberharceleur.

   Une deuxième méthode consiste à fournir une sorte d'éducation interactive. À l'école, les campagnes de sensibilisation n'ont pas d'effets directs : les élèves ne sont pas toujours conscients de la portée du cyberharcèlement. Il est possible que les jeunes aient une autre interprétation du phénomène et qu'ils ne se sentent pas visés. L'éducation interactive veut d'un côté inciter le cyberharceleur à réfléchir sur l'action qu'il est sur le point de faire en montrant par exemple des vidéos dans lesquelles des victimes de cyberharcèlement témoignent de ce qui leur est arrivé. D'un autre côté, l'éducation interactive porte secours aux victimes en leur suggérant la manière dont elles peuvent réagir à un message d'un cyberharceleur ou en les renvoyant à des organisations spécialisées.
2) **Retard**. Parfois un cyberharceleur regrette qu'il a posté un message tout de suite après l'action[25]. En donnant un utilisateur un peu de temps avant qu'il n'envoie un message, on peut éviter bon nombre de cas de cyberharcèlement. Si le système ne publie le message qu'après un certain délai de temps, il offre à l'utilisateur du temps à réfléchir et la possibilité d'annuler l'envoi.
3) **Conséquences**. Le manque de contact réel fait que les cyberharceleurs perdent de vue les conséquences éventuelles qui suivent leur action. Il serait utile de montrer quels sont les effets de

[25] Voir à ce sujet: Wang, Yang e.a. 2011; Xu, Jun-Ming e.a. 2013.

poster un message : le système pourrait par exemple dire combien de personnes liraient le message, pour que le cyberharceleur se rende compte de la portée de son action.

4) **Matière éducationnelle**. Outre que l'éducation interactive, il peut y avoir aussi une éducation ciblée. La matière éducationnelle offre surtout de l'aide aux victimes sous forme des soi-disant stratégies d'adaptation (*coping strategies*) : la victime reçoit des informations qui l'aident à relativiser le message. À partir du contenu d'un message blessant, le système pourrait recommander quelques vidéos qui portent sur la même thématique. Supposons qu'un commentaire critique l'orientation sexuelle de la victime, celle-ci serait dirigée alors à des vidéos d'une organisation contre l'homophobie.
5) **Signalement**. La dernière possibilité est la plus facile. Nous avons déjà mentionné que peu d'utilisateurs des médias sociaux signalent les messages inappropriés. Le système pourrait proposer de signaler un commentaire douteux aux modérateurs, de sorte qu'ils y donnent la priorité.

Les trois premiers modes d'intervention visent à éviter que le cyberharceleur poste son message. Ils partent de l'idée que le comportement doit changer pour diminuer les cas de cyberharcèlement. Au moyen d'une intervention directe, on espère que le cyberharceleur se rendra compte des effets de son action. Les deux derniers servent plutôt à mitiger directement après que quelqu'un a reçu un message blessant.

Notons que les interventions mentionnées ci-dessus ont lieu dans un seul réseau social. Un cyberharceleur persistent pourrait facilement continuer ses actions dans une autre communauté sociale. Voilà pourquoi Dadvar e.a. (2012) proposent une approche transmédiale. De nos jours, beaucoup de gens ont plusieurs comptes sur divers réseaux sociaux. Un compte *YouTube* est lié automatiquement à un compte *Google+*, qu'on peut à son tour lier à *Twitter*, *Facebook* et tant d'autres médias sociaux. Les chercheurs pensent que bon nombre de cyberharceleurs peuvent être arrêtés en suivant les traces qu'ils ont laissées en ligne. À l'aide des données métatextuelles, il serait possible de suivre l'utilisateur pendant une période plus longue. Cela permettrait aussi aux modérateurs de voir si l'utilisateur est vraiment un cyberharceleur (par le comportement répétitif) ou plutôt une victime (qui a tapé pour une seule fois une réponse agressive). De telle façon, le cyberharceleur pourrait être banni de plusieurs sites internet à la fois.

Ce parcours limité a donné plusieurs idées sur différents modes d'intervention : ou bien ils sont remédiants, ou bien mitigeants, quoique l'on préfère la première solution. Selon nous, l'éducation interactive et la visualisation des conséquences sont des pistes prometteuses, parce qu'elles agissent sur le comportement de l'individu. En plus, pour trouver les cyberharceleurs persistants dans la communauté, il vaut la peine d'examiner tous les comptes en ligne des utilisateurs suspects.

4. Conclusions générales et pistes de recherche à suivre dans l'avenir

Dans ce mémoire, nous espérons avoir montré l'importance d'une détection (semi-) automatique de cas de cyberharcèlement. Au premier chapitre, nous avons esquissé la portée de la problématique. Les dernières années, le cyberharcèlement est devenu plus fréquent et plus urgent. Les millions de commentaires postés chaque jour, compliquent énormément la tâche des modérateurs des réseaux sociaux. Il est donc souhaitable que la linguistique informatique essaye d'automatiser la détection de messages offensifs. Nous présenterons ici nos conclusions, qui répondent aux questions de recherche formulées dans l'introduction.

1) *Quels traits linguistiques caractérisent les messages des victimes de cyberharcèlement et des cyberharceleurs dans les médias sociaux?*

Commençons par dire qu'il y a en effet des traits typiques qui indiquent le cyberharcèlement, et par conséquent il est possible d'automatiser la détection de cas de cyberharcèlement textuel. Nous avons remarqué plusieurs traits appartenant à divers niveaux linguistiques. Ainsi, nous avons constaté que la présence des injures et de certains mots clés est fort pertinente. En plus, les messages ont dans la plupart des cas un contenu négatif (voire sarcastique). Pour finir, nous avons vu que le thème d'une conversation peut changer tout à coup et qu'un message négatif 'attire' en quelque sorte d'autres énoncés dépréciatifs .

2) *Quelles techniques de la linguistique informatique permettent de détecter les traits dégagés des messages de cyberharcèlement ?*

Après notre analyse d'un cas de cyberharcèlement, nous avons parcouru systématiquement les techniques computationnelles qui permettent la détection des traits linguistiques dégagés. Nous avons proposé (à la fin du deuxième chapitre) un schéma qui regroupe les techniques que nous retenons comme les plus prometteuses. Selon nous, le calcul de la valence d'un message (par l'analyse sentimentale, les émoticônes et la base de données *ConceptNet*) serait un bon point de départ. Ensuite, le système doit tenir compte du contexte en comparant le thème d'un message initial avec celui du commentaire concerné. N'oublions pas non plus le principe de la contextualité : ce qui se ressemble s'assemble (un commentaire négatif en attire d'autres). Enfin, une liste de mots clés (parmi lesquels se trouvent également les injures) permet de classer les cas de cyberharcèlement par degré d'urgence. De telle manière, les modérateurs voient du premier coup d'œil tous les cas dangereux, ce qui leur permet de prendre des mesures efficaces.

3) *Comment peut-on résoudre ou éviter que le cyberharcèlement se poursuive après la détection automatique?*

Le troisième chapitre a porté entièrement sur les différents modes d'intervention après la constatation d'un cas présumé de cyberharcèlement. Ces modes sont soit directs, soit remis. Nous entendons par intervention directe : « empêcher que l'utilisateur envoie un message de cyberharcèlement ». Il y a trois modes d'intervention directs : une interruption de l'action, la possibilité de remettre l'envoi d'un message ou une

explication des conséquences du cyberharcèlement. Quand un message a été envoyé, une aide immédiate doit limiter les dommages psychologiques. L'éducation au moyen de vidéos ou de sites internet et le signalement de messages blessants s'avèrent de bonnes méthodes pour aider les victimes. Cette pratique veut rendre conscients les utilisateurs de leur conduite en ligne. Contrairement à l'école, une intervention automatique montre au moment nécessaire quand quelqu'un se comporte de façon inacceptable.

4.1 Pistes de recherche à suivre dans l'avenir

Dorénavant il est aux chercheurs de développer un système qui fonctionne bien, de sorte que les enfants puissent jouir d'un monde virtuel plus sûr. Il ne faut cependant pas simplement censurer tous les messages blessants ; il est important que les internautes se rendent compte de leur comportement transgressif. Comment peut-on améliorer la détection de cas de cyberharcèlement et quelles sont les pistes à suivre?

D'abord, il faudra que les techniques soient mises au point. En construisant un corpus annoté et en le rendant disponible à tout le monde, on constatera probablement des progrès plus rapides. Il est également nécessaire d'élaborer des listes d'injures, de mots clés et de relations stéréotypées pour augmenter les performances du système.

Une approche transmédiale est une idée que nous avons évoquée brièvement. Elle consiste en la mise en relation de plusieurs réseaux sociaux, de sorte que les divers comptes d'un même utilisateur soient unis. De telle manière, le système est valable pour plusieurs communautés en ligne et les modérateurs des différents réseaux sociaux savent quels utilisateurs sont victimes et lesquels sont des cyberharceleurs persistents.

Dans notre mémoire, nous ne nous sommes occupés que de messages publics postés et partagés consciemment dans les médias sociaux. Pour garantir une détection profonde, les messages privés doivent aussi être pris en considération. Afin de ne pas violer la vie privée d'un utilisateur, il est nécessaire que le système détecte automatiquement (c'est-à-dire sans l'intervention d'un modérateur) les messages de cyberharcèlement. Pour y parvenir, la précision devra augmenter considérablement.

Enfin, on peut élargir la détection de cas de cyberharcèlement à la détection d'images choquantes, éventuellement accompagnées d'une description textuelle. Nous avons parlé de la nature transmédiale (l'ordinateur, l'ordiphone, le portable…) et transmodale (photos, texte, vidéos…) du cyberharcèlement. La reconnaissance optique de caractères (ROC) permet aux systèmes informatiques de 'lire' un texte sur une photo ou une vidéo et offre de bonnes perspectives pour l'avenir. Les chercheurs peuvent donc essayer de trouver une solution à ces cas non-textuels.

5. Références

Bayzick, Jennifer, Kontostathis, April & Edwards, Lynne, 'Detecting the Presence of Cyberbullying Using Computer Software.' In: *Proceedings of the ACM WebSci'11*, Koblenz, 2011, 1 – 2.

Campbell, Marilyn A., 'Cyberbullying: an old problem in a new guise?', *Australian Journal of Guidance and Counselling* 15 (2005), 68 – 76.

Chisholm, June F., 'Cyberspace Violence against Girlsand Adolescent Females', *Annals of the New York Academy of Sciences* 1087 (2006), 74 – 89.

Dadvar, Maral & de Jong, Franciska, 'Cyberbullying Detection: A Step Toward a Safer Internet Yard.' In: Mille, Alain e.a. (éd.), *Proceedings of the 21st World Wide Web Conference 2012 (16 – 20 April 2012)*, Lyon, 2012, 121 – 125.

Dadvar, Maral e.a., 'Improved Cyberbullying Detection Using Gender Information.' In: *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (23-24 Feb 2012)*, Ghand, 2012, 23 – 25.

Dadvar, Maral e.a., 'Towards User Modelling in the Combat against Cyberbullying.' In: Bouma, Gosse e.a. (éd.), *17th International Conference on Applications of Natural Language to Information Systems*, NLDB (26-28 Juin 2012), Groningue, 2012, 277 – 283.

Dinakar, Karthik e.a., 'Commonsense Reasoning for Detection, Prevention and Mitigation of Cyberbullying', *ACM Transactions on Interactive Intelligent Systems* 2 (2012) 3, doi: 10.1145/2362394.2362400 [8/10/2012].

González-Ibáñez, Roberto, Muresan, Smaranda & Wacholder, Nina, 'Identifying Sarcasm in Twitter: A Closer Look.' In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers*, Oregon, 2011, 581 – 586.

Greene, Michael B., 'Bullying and harassment in schools.' In: Moser, Rosemarie Scolaro & Frantz, Corinne E. (éds), *Shocking violence: Youth perpetrators and victims – A multidisciplinary perspective,* Springfield, 2000, 72 – 101.

Harris, Sandra & Petrie, Garth, 'A study of bullying in the middle school', *National Association of Secondary School Principals (NASSP) Bulletin* 86 (2002), 42 – 53.

Herring, Susan C., *Gender differences in computer-mediated communication: bringing familiar baggage to the new frontier*. http://mith.umd.edu//WomensStudies/Computing/Articles+ResearchPapers/gender-differences-communication [19/09/2012].

Kontostathis, April, Edwards, Lynne & Leatherman, Amanda, 'Text Mining and Cybercrime.' In: Berry, Michael W. & Kogan, Jacob (éds), *Text Mining: Applications and Theory*, Chichester, 2010, 149 – 164.

Langos, Colette, 'Cyberbullying: the Challenge to Define', *Cyberpsychology, Behavior and Social Networking* 15 (2012), 285 – 289.

Li, Qing, 'Cyberbullying in Schools: a Research of Gender Differences', *School Psychology International* 27 (2006), 157 – 170.

Liu, Hugo & Singh, Push, 'ConceptNet – a practical commonsense reasoning tool-kit', *BT Technology Journal* 22 (2004), 211 – 226.

Omernick, Eli & Sood, Sara Owsley, 'The Impact of Anonymity in Online Communities', *International Conference on Intelligent User Interfaces*, mars 2013 [pas encore publié].

Polanyi, Livia & Zaenen, Annie, 'Contextual Valence Shifters.' In: Shanahan, James G. e.a. (éds), *Computing Attitude and Affect in Text: Theory and Applications*, Pays-Bas, 2006, 1 – 10.

Ptaszynski, Michal e.a., 'In the Service of Online Order: Tackling Cyber-bullying with Machine Learning and Affect Analysis.' In: *International Journal of Computational Linguistics Research* 1 (2010*)*, 135 – 154.

Reynolds, Kelly, Kontostathis, April & Edwards, Lynne, 'Using Machine Learning to Detect Cyberbullying.' In: *Proceedings of the 2011 10th International Conference on Machine Learning and Applications Workshops* (ICMLA 2011), Honolulu, 2011, 241 – 245.

Ritter, Alan e.a., 'Named Entity Recognition in Tweets: An Experimental Study.' In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (CEMNLP), Edinburgh, 2011.

Salmivalli, Christina, 'Participant role approach to school bullying: implications for interventions', *Journal of Adolescence* 22 (1999), 453 – 459.

Slonje, Robert & Smith, Peter K., 'Cyberbullying: another main type of bullying?', *Scandinavian Journal of Psychology* 49 (2008), 147 – 154.

Sood, Sara Owsley, Antin, Judd & Churchill, Elizabeth F., 'Profanity use in online communities.' In: *Proceedings of ACM SIGCHI*, Austin, 2012a.

Sood, Sara Owsley, Antin, Judd & Churchill, Elizabeth F., 'Automatic Identification of Personal Insults on Social News Sites', *Journal of the American Society for Information Science and Technology* 63 (2012b), 270 – 285.

Vandebosch, Heidi & Van Cleemput, Katrien, 'Cyberbullying among youngsters: profiles of bullies and victims', *New Media and Society* 11 (2009), 1349 – 1371.

Vandebosch, Heidi e.a., *Cyberpesten bij jongeren in Vlaanderen*: *studie in opdracht van het viWTA*, Brussel, 2006, [consultable en ligne sur:] http://www.samenlevingentechnologie.be/ists/nl/pdf/rapporten/rapportcyberpesten.pdf [30/01/2013].

Wang, Yang e.a., 'I regretted the minute I pressed schare : a qualitative study of regrets on Facebook.' In: *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, New York, 2011, 1 – 16.

Xu, Jun-Ming, Zhu, Xiaojin & Bellmore, Amy, 'Fast Learning for Sentiment Analysis on Bullying.' In: *ACM KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM),* Beijing, 2012a.

Xu, Jun-Ming e.a., 'Learning from Bullying Traces in Social Media.' In: Fosler-Lussier, Eric e.a. (éds), *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, 2012b, 656 – 666.

Xu, Jun-Ming e.a., 'An Examination of Regret in Bullying Tweets.' In: *Proceedings of the 2013 Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL HLT), Atlanta, 2013.

Ybarra, Michele L. & Mitchell, Kimberly J., 'Online Aggressor/Targets, Aggressors, and Targets: A Comparison of Associated Youth Characteristics', *Journal of Child Psychology and Psychiatry* 45 (2004), 1308 – 1316.

Yin, Dawei e.a., 'Detection of Harassment on Web 2.0.' In: *Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0)*, Madrid, 2009.