3 ème édition des journées scientifiques du Laboratoire d'Informatique et d'Ingénierie pour l'Innovation

Du 23 au 25 février 2023 - Université Assane Seck de Ziguinchor – Cap Skiring

Ibrahima NDAO, Université Assane SECK de Ziguinchor - Sénégal, LI3
i.ndao20150570@zig.univ.sn

UFR Sciences et Technologies
Laboratoire d'Informatique et d'Ingénierie pour l'Innovation

Thème : Traitement des phénomènes linguistiques pour la détection des messages à caractères abusifs

Résumé :

L'utilisation des plateformes numériques est devenue une pratique quotidienne. Cependant, la popularité de ces plateformes et leurs libertés d'usage les confrontent à un nombre croissant d'utilisateurs aux comportements abusifs. Cela a suscité l'attention des gouvernements qui exigent aux propriétaires de ces plateformes d'accroître leurs efforts pour lutter contre ce phénomène, à travers différents lois et règlements. Le langage abusif renferme l'ensemble des communications offensantes, intimidantes, fausses, exagérées ou d'attaques envers une personne ou une communauté sur la base de certaines caractéristiques telles que la couleur, l'origine, l'ethnie, l'orientation sexuelle, etc. Plusieurs méthodes et techniques ont été mises en œuvre pour lutter contre ce phénomène. Toutefois, des techniques de contournement telles que le camouflage de messages, l'utilisation d'abréviations, la saisie phonétique, le mélange de codes (code mixing en anglais) sont de plus en plus utilisés. D'où le besoin impératif d'une solution plus efficace permettant de contrôler ces types de messages. L'objectif de notre travail est de proposer une approche de détection des messages à caractères abusifs. Nous avons fait une revue des travaux connexes sur la détection des messages à caractères abusifs.

Mots clés : message à caractères abusifs, phénomènes linguistiques, médias sociaux, traitement automatique du langage naturel (TALN), fouille de textes.

Etat de l'art

Ces dernières années, l'utilisation massive des médias sociaux et la liberté d'expression qui va avec, favorise de plus en plus de comportements abusifs. Le langage abusif renferme l'ensemble des communications offensantes, intimidantes, fausses, exagérées ou d'attaques envers une personne ou une communauté sur la base de certaines caractéristiques telles que la couleur, l'origine, l'ethnie, l'orientation sexuelle, etc. Ceci fait que les recherches sur la détection des messages à caractère abusif sont en nette croissance. Plusieurs chercheurs ont proposé des solutions, classant les commentaires (tweets) dans différentes catégories. Les classes de messages abusifs les plus étudiées sont les messages haineux, les messages offensants, le cyber harcèlement, les messages racistes, sexistes, le trolling, etc. Les travaux antérieurs démarrent soit par la constitution de corpus spécialisé ou de l'utilisation de corpus existant; la plupart de ces études se sont concentrées sur l'anglais et l'arabe. Les premières études sur la détection de message abusif reposent sur l'approche non supervisée en exploitant des lexiques de mots ; quant à la plupart des travaux se basent sur une approche supervisée et récemment sur l'apprentissage profond. Plusieurs techniques de classification ont été utilisées notamment les naïve bayes, les SVM, la régression logistique, les CNN, les RNN, le GRU, les LSTM, etc. Ils se sont intéressées à différentes caractéristiques soit au niveau des mots, soit au niveau des séquences des caractères/mots ou bien même au niveau des mots intégrés. La multi pluralité de ces recherches montre la complexité de ce champ d'étude tant dans la classification spécifique des tweets que des phénomènes linguistiques subtiles.

Domaine d'applications

construction chatterbots

recommandation de contenu

l'extraction d'événements controverses

l'analyse de sentiments

Bibliographies

R. Toujani, « Opinions Mining from Posters' Users in Social Networks ».
N. Cécillon, R. Dufour, et V. Labatut, « Approche multimodale par plongement de texte et de graphes pour la détection de messages abusifs », p. 26.
H. Mubarak, K. Darwish, et W. Magdy, « Abusive Language Detection on Arabic Social Media », in *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada, août 2017, p. 52-56. doi: 10.18653/v1/W17-3008.
T. Davidson, D. Warmsley, M. Macy, et I. Weber, « Automated Hate Speech Detection and the Problem of Offensive Language ». arXiv, 11 mars 2017. Consulté le: 22 novembre 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1703.04009
T. Davidson, D. Warmsley, M. Macy, et I. Weber, « Automated Hate Speech Detection and the Problem of Offensive Language ». arXiv, 11 mars 2017. Consulté le: 22 novembre 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1703.04009
V. D. Gawali, T. D. Bhalerao, S. G. Kamble, R. Basak, et S. Sural, « Abusive language Detection using Machine Learning », vol. 7, nº 5, p. 6, 2020.
I. Kwok et Y. Wang, « Locate the Hate: Detecting Tweets against Blacks », *Proc. AAAI Conf. Artif. Intell.*, vol. 27, nº 1, p. 1621-1622, juin 2013, doi: 10.1609/aaai.v27i1.8539.