Détection de tweets offensif en contexte

Encadrement : ygor.gallina@univ-nantes.fr
Lieu de TER : LS2N (Site de la FST)
Nombre d'étudiant·es : 1/2 avec une appétence pour le TAL,

Contexte

Ces dernières années ont été marquées par la prolifération de messages offensant sur les médias sociaux tels que Facebook ou Twitter. Pour protéger les utilisateur·ices les messages peuvent être signaler par d'autres usager·es de la plateforme, les messages peuvent ensuite être filtrés manuellement par l'entreprise. Ce filtrage manuel est chronophage et peut provoquer des symptômes de stress post-traumatique chez les annotateurs humains c'est pourquoi un pan de la recherche en TALN s'intéresse à automatiser le processus.

La plupart des travaux sur la détection de messages offensifs se concentrent sur le contenu du message sans tenir compte du contexte du message. Ce contexte est primordial pour s'assurer que le message n'est pas un faux positif. Suivant le contexte un message contenant une insulte par exemple peut ou ne pas être considéré comme offensif.

Sujet

L'objectif de ce travail est de s'intéresser à l'historique de discussion d'un message pour comprendre s'il est offensif ou non.

- Identifier les tweets et traits pertinent pour la détection de tweets offensif dans l'historique de discussion (fils twitter)
- Choisir et extraire les traits identifiés pour entraîner un classifieur de tweets offensifs

References

[1] M. Dadvar, D. Trieschnigg, R. Ordelman, et F. de Jong, « Improving cyberbullying detection with user context », in Proceedings of the 35th European conference on Advances in Information Retrieval, Berlin, Heidelberg, mars 2013, p. 693-696. doi: 10.1007/978-3-642-36973-5_62.

[2] P. Fortuna, J. Soler, et L. Wanner, « Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets », in Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, mai 2020, p. 6786-6794. https://aclanthology.org/2020.lrec-1.838

[3] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, et M. Granitzer, « I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language », in Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, mai 2020, p. 6193-6202. https://aclanthology.org/2020.lrec-1.760

[4] M. Diaz, R. Amironesei, L. Weidinger, et I. Gabriel, « Accounting for Offensive Speech as a Practice of Resistance », in Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Seattle, Washington (Hybrid), juill. 2022, p. 192-202. doi: 10.18653/v1/2022.woah-1.18.