Exploration de caractéristiques d'embeddings de graphes pour la détection de messages abusifs

Mots-clés : classification automatique, traitement automatique du langage, *embeddings* de graphes, messages abusifs.

Contexte

Les enjeux scientifiques et sociétaux liés à la surveillance des échanges transitant sur la toile sont considérables, en particulier parce que l'internet « interactif » s'est développé massivement ces dernières années et qu'il est devenu le support privilégié d'échanges libres entre des individus potentiellement très différents. Le bon fonctionnement des espaces d'échange ouverts implique un contrôle des interactions qui est souvent difficile[1,2]. En effet, l'absence de modération peut conduire à un détournement du dispositif ou au développement de comportements déviants, illégaux ou, plus généralement, nuisibles à la qualité des échanges (exemple des *trolls* sur Internet). Si l'absence de contrôle est un des facteurs qui a contribué au développement des espaces d'échanges ouverts sur la toile, l'absence de modération peut aussi en limiter l'intérêt.

Lorsque des utilisateurs enfreignent les règles de la communauté, des sanctions peuvent alors être appliquées. Ce processus, appelé *Modération*, est principalement effectué par des opérateurs humains. Ce travail manuel étant coûteux, les entreprises ont tout intérêt à automatiser le processus. Nous considérons le problème de classification consistant à déterminer automatiquement si un message posté dans une conversation sur Internet est abusif ou non.

Les approches pour détecter automatiquement les messages abusifs s'appuient généralement sur le contenu des messages ciblés uniquement, comme par exemple la définition de règles statiques pour extraire des marqueurs linguistiques [Spartus97], l'utilisation de caractéristiques syntaxiques et lexicales [Chen12], ou encore des approches évoluées d'apprentissage automatique [Chavan15]. Certaines œuvres hybrides proposent également de combiner les deux catégories. Cependant, ces méthodes doivent faire face à différents problèmes : les abus peuvent être répartis sur une succession de messages, être masqués volontairement par les utilisateurs pour éviter la détection de mots incorrects (par exemple modifier intentionnellement l'orthographe d'un mot interdit [Hosseini17]), dépendre d'un contexte plus large que celui de la conversation considérée…

Ainsi, d'autres travaux ont décidé de considérer également le contenu des messages autour du message ciblé, que l'on cherche à identifier comme abusif ou non, au lieu de considérer le contenu du message ciblé uniquement [Yin09]. De même, des approches utilisant des méta-informations sur les utilisateurs [Balci15] ou des modèles d'utilisateurs [Cheng15] ont également été explorées.

Description du stage

Nous avons proposé deux méthodes différentes pour classer automatiquement les messages abusifs :

- L'une est basée sur des caractéristiques textuelles classiques basées sur le message lui-même et la conversation environnante enrichie d'approches de prétraitement de texte originales [Papegnies17a] ;

1 http://www.slate.fr/story/88227/commentaires-articles-ruinent-medias

2 https://scholarworks.iu.edu/dspace/handle/2022/1020

- L'autre ignore complètement le contenu des messages et modélise les conversations sous forme de graphes de conversation [Papegnies17b].

Les premiers résultats obtenus sur un corpus de conversations textuelles d'une messagerie instantanée ont montré que l'approche utilisant des caractéristiques issues de graphes modélisant la structure et la dynamique de conversations (et donc ignorant le contenu du message), permettent d'obtenir de meilleures performances que les approches basiques de traitement automatique du langage, s'appuyant quant à elles sur le contenu textuel seul des messages échangés. **Un des premiers enjeux de ce stage** pourra être de reprendre ce travail de recherche pour explorer la combinaison des caractéristiques textuelles et des caractéristiques des graphes de conversation.

L'approche par graphes de conversation [Papegnies17b] extrait, comme caractéristiques, un certain nombre de mesures topologiques classiques, qui sont ensuite dans un processus de classification. Cependant, l'extraction de caractéristiques de ces graphe peut être un processus long et coûteux en termes de calcul. Récemment, des méthode d'*embeddings* de graphes ont été proposées pour pallier ces difficultés : les données du graphe sont converties en un espace de petite dimension dans lequel les informations de structure et les propriétés du graphe sont préservées au maximum [Cai18]. **Le premier objectif** sera de faire un état de l'art des approches par *embeddings* de graphes existantes, en les appliquant sur ce problème de détection de messages abusifs. **Le second objectif**, plus exploratoire, sera de proposer une modélisation des graphes de conversation en utilisant à la fois la structure de la conversation, mais également le contenu textuel de celles-ci, qui est pour l'instant ignoré dans la modélisation que nous proposons. Des *embeddings* de ces graphes pourront ensuite être utilisés et comparés aux anciens sur cette tâche de classification.

Encadrants

Vincent Labatut – Maître de conférence – vincent.labatut@univ-avignon.fr
Richard Dufour – Maître de conférence – richard.dufour@univ-avignon.fr

Références

[Balci15] Balci K., Salah A. A. (2015). *Automatic analysis and identification of verbal aggression and abusive behaviors for online social games*. Computers in Human Behavior, 53, 517-526.

[Chavan15] Chavan V. S., Shylaja S. S. (2015). *Machine learning approach for detection of cyber-aggressive comments by peers on social media network*. In : Advances in computing, communications and informatics (ICACCI), pp. 2354-2358. IEEE.

[Cai 18] Cai H., Zheng V. W., Chang K. (2018). *A comprehensive survey of graph embedding: problems, techniques and applications*. IEEE Transactions on Knowledge and Data Engineering.

[Chen12] Chen Y., Zhou Y., Zhu S., Xu H. (2012). *Detecting offensive language in social media to protect adolescent online safety*. In : Privacy, Security, Risk and Trust (PASSAT), pp. 71-80.

[Cheng15] Cheng J., Danescu-Niculescu-Mizil C., Leskovec J. (2015). *Antisocial Behavior in Online Discussion Communities*. In Icwsm, pp. 61-70.

[Hoseini17] Hosseini H., Kannan S., Zhang B., Poovendran R. (2017). *Deceiving Google's Perspective API Built for Detecting Toxic Comments*. arXiv preprint arXiv:1702.08138.

[Papegnies17a] Papegnies E., Labatut V., Dufour R., Linares G. (2017). *Impact of content features for automatic online abuse detection*. In : CORIA 2017.

[Papegnies17b] Papegnies E., Labatut V., Dufour R., Linares G. (2017). *Graph-based Features for Automatic Online Abuse Detection*. In : International Conference on Statistical Language and Speech Processing, pp. 70-81.

[Spertus97] Spertus E. (1997). *Smokey: Automatic recognition of hostile messages*. In : 14th National Conference on Artificial Intelligence and 9th Conference on Innovative Applications of Artificial Intelligence, AAAI, 1997, pp. 1058–1065.

[Yin09] Yin D., Xue Z., Hong L., Davison B. D., Kontostathis A., Edwards L. (2009). *Detection of harassment on web 2.0*. Proceedings of the Content Analysis in the WEB, 2, 1-7.