Détection de la controverse : une approche basée sur les réseaux de neurones, appliquée aux graphes et aux textes

S. Benslimane[1], J. Aze[1], S. Bringay[2,1], C. Mollevi[3,4], M. Servajean[2,1]

[1] LIRMM UMR 5506, CNRS, University of Montpellier, Montpellier, France
[2] AMIS, Paul Valery University, Montpellier, France
[3] Institut du Cancer Montpellier (ICM), Montpellier, France
[4] Institut Desbrest d'Epidémiologie et de Santé Publique, UMR Inserm

prenom.nom@lirmm.fr

Résumé

Cet article propose une approche de détection de la controverse dans les réseaux sociaux, basée sur la structure d'une discussion et de ses caractéristiques textuelles. La méthode proposée s'appuie sur les réseaux de neurones graphiques (Graph Neural Networks ou GNN) pour encoder la représentation graphique de la discussion (y compris les textes) dans un vecteur multidimensionnel. Ce dernier est utilisé pour classer les fils de discussions comme étant controversés ou non. Les expériences menées sur différents jeux de données montrent l'impact positif de la combinaison des caractéristiques textuelles et structurelles.

Mots-clés

Détection de controverse, Réseaux de neurones graphiques, Traitement naturel du langage, Reddit

Abstract

This paper proposes a controversy detection approach based on both graph structure of a discussion and text features. Our proposed approach relies on Graph Neural Network (GNN) to encode the graph representation (including its texts) in an embedding vector before performing a graph classification task. The latter will classify the post as controversial or not. Conducted experiments using different real-world datasets show the positive impact of combining textual features and structural information.

Keywords

Controversy detection, Graph neural networks, Natural language processing, Reddit

1 Introduction

Cet article [1] est un résumé de la publication réalisée pour la conférence WISE 2021 [1] et étendue dans le journal World Wide Web [2]. La disponibilité d'un grand nombre de sources de données et l'émergence de réseaux sociaux, tels que Twitter et Reddit, ont accru la connectivité sociale des personnes, ce qui leur permet d'exprimer, de propager, de partager et de contester facilement des opinions. Dans cet article, nous étudions le phénomène social de la controverse. Un contenu controversé peut être défini comme un contenu attirant des avis et opinions divergents, tant positifs que négatifs [4]. La détection précoce de ces sujets est importante, pour éviter par exemple la désinformation ou les discussions haineuses. Cependant, il s'agit d'une tâche difficile qui doit être effectuée sur un grand nombre de contenus, en constante évolution et couvrant un large éventail de thématiques. La controverse évolue au cours du temps et selon les communautés engagées.

Pour résoudre cette tâche, on trouve dans la littérature trois types de travaux : **(i)** les approches basées sur le contenu, **(ii)** celles basées sur la structure et **(iii)** celles considérées comme hybrides. Les premières utilisent uniquement les caractéristiques textuelles des messages [5]. Cependant, l'interprétation des messages et des termes utilisés étant subjectif, l'information comprise dans ces textes peut être différente selon certains facteurs, tels que la culture ou la langue des communautés, et doit donc être traitée avec précaution. Le second type d'approche se base sur les interactions entre utilisateurs, révélées par des informations structurelles issues du graphe des interactions de ces utilisateurs (e.g. propriété de connectivité ou de centralité) [3]. Enfin, des études récentes combinent les informations issues du contenu et de la structure [9].

Nous présentons dans cet article une nouvelle approche hybride de détection de la controverse, basée sur les réseaux de neurones graphiques (Graph Neural Networks) afin de combiner les informations textuelles et structurelles. L'originalité de notre approche réside dans l'utilisation de méthodes GNNs pour représenter les utilisateurs (nœuds) dans un espace euclidien à faible dimension, en tenant compte des informations structurelles. Les deux architectures proposées, l'une basée sur une représentation hiérarchique du graphe, l'autre sur des mécanismes d'attention, diffèrent largement de l'approche de [10]. Nos expérimentations se focalisent sur des données réelles du réseau social Reddit, même si notre méthode est applicable à tout autre média social suivant quelques adaptations lors de la construction du graphe.

1. L'article a reçu le prix du meilleur article de la conférence WISE.

2 Méthode

Notre approche se décompose en 4 étapes :
Étape 1 : Construction du graphe. Un fil de discussion est représenté sous la forme d'un graphe non orienté où un nœud représente un utilisateur et une arête entre 2 nœuds représente une réponse d'une personne à une autre.
Étape 2 : Caractéristiques des utilisateurs. Chaque utilisateur est représenté par les contenus qu'il a publiés dans la discussion. Récemment, différents modèles de langage NLP tels que BERT, pré-entraînés sur un large corpus, ont été proposés pour améliorer la représentation dynamique du texte. Nous extrayons, pour chaque texte, un vecteur le représentant à partir du modèle BERT, et nous agrégeons ensuite ces vecteurs par utilisateur.
Étape 3 : Encodage du graphe. Cette étape vise à représenter l'ensemble du graphe sous la forme d'un vecteur à faible dimension. Ce dernier sera utilisé en entrée de la dernière étape de classification du graphe. Récemment, différentes approches basées sur les GNNs ont été proposées pour adapter les architectures d'apprentissage profond aux données de type graphe [6, 7]. L'idée principale est de considérer chaque nœud du graphe comme un nœud de calcul, et d'apprendre à partir des GNNs un plongement dans un espace vectoriel représentant les nœuds. Cette étape exploite à la fois les caractéristiques des nœuds de la couche précédente ainsi que la représentation de ces voisins. Ensuite, la représentation de ces nœuds est agrégée afin de représenter le graphe complet. Deux stratégies sont proposées pour cette représentation vectorielle du graphe : la première basée sur les représentations hiérarchiques d'un graphe par des réseaux convolutifs [8] et la seconde basée sur des scores d'attention entre les nœuds [7].
Étape 4 : Classification du graphe. À l'aide du vecteur représentant le graphe et d'un réseau de neurones, le fil de discussion est ensuite classifié, controversé ou non.

3 Expériences

Les expériences ont été menées sur plusieurs jeux de données (subreddits) provenant de Reddit [4], chacun de ces jeux comprenant des milliers de fils de discussions, composés de leurs messages respectifs. En comparant nos résultats à des méthodes utilisant ces données et se basant soit seulement sur la structure [4], soit sur la structure combinée à des informations textuelles [4, 10], les deux méthodes proposées obtiennent des résultats équivalents en termes de précision de classification des fils de discussions controversés ou non, voire supérieurs sur certains jeux de données, notamment pour notre approche basée sur la représentation hiérarchique du graphe. Des expériences, omettant les informations textuelles dans le graphe, ont montré que la précision de la classification dans certains jeux de données donnaient de moins bons résultats.

4 Conclusion

Nous avons présenté dans ce résumé nos travaux autour de la détection de la controverse sur Reddit, combinant les informations structurelles et textuelles autour de l'utilisation des réseaux de neurones graphiques (GNNs). Nous prévoyons d'étendre ces travaux pour quantifier la controverse et prendre en compte la temporalité.

Références

[1] Samy Benslimane, Jérôme Azé, Sandra Bringay, Maximilien Servajean, and Caroline Mollevi. Controversy detection : a text and graph neural network based approach. In *Web Information Systems Engineering*, 2021.

[2] Samy Benslimane, Jérôme Azé, Sandra Bringay, Maximilien Servajean, and Caroline Mollevi. A text and GNN based controversy detection method on social media. *World Wide Web*, 26(2) :799–825, 2023.

[3] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1) :3 :1–3 :27, 2018.

[4] Jack Hessel and Lillian Lee. Something's brewing ! early prediction of controversy-causing posts from discussion features. In *ACL Human Language Technologies, Volume 1*, pages 1648–1659, 2019.

[5] Myungha Jang, John Foley, Shiri Dori-Hacohen, and James Allan. Probabilistic approaches to controversy detection. In *25th ACM International Conference on Information and Knowledge Management, CIKM*, pages 2069–2072, 2016.

[6] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Int. Conf. on Learning Representations*, 2017.

[7] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Int. Conf. on Learning Representations*, 2018.

[8] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, pages 4805–4815, 2018.

[9] Juan Manuel Ortiz De Zarate and Esteban Feuerstein. Vocabulary-based method for quantifying controversy in social media. In *Int. Conf. on Conceptual Structures*, volume 12277, pages 161–176, 2020.

[10] Lei Zhong, Juan Cao, Qiang Sheng, Junbo Guo, and Ziang Wang. Integrating semantic and structural information with graph convolutional network for controversy detection. In *Proceedings of the 58th Annual Meeting of ACL*, pages 515–526, July 2020.

Remerciements

Ce projet a été soutenu par des subventions du fond de dotation Janssen Horizon. L'accès aux ressources HPC de l'IDRIS a été accordé dans le cadre de l'allocation AD011012604 par GENCI.