Analyse, détection et quantification de la controverse dans les médias sociaux par l'utilisation de modèle d'apprentissage profond sur des données structurelles et textuelles

Samy Benslimane

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale : Information, Structures, Systèmes

Unité de recherche : Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), France

Analyse, détection et quantification de la controverse dans les médias sociaux par l'utilisation de modèles d'apprentissage profond sur des données structurelles et textuelles

Présentée par Samy BENSLIMANE
Le 8 Décembre 2023

Sous la direction de Sandra BRINGAY, Caroline MOLLEVI et Maximilien SERVAJEAN

Devant le jury composé de

Diana INKPEN, Professeur, University of Ottawa	Rapporteure
Serena VILLATA, Directrice de recherche, CNRS - Université Côted'Azur	Rapporteure
Engelbert MEPHU NGUIFO, Professeur des universités, Université Blaise Pascal Clermont	Président
Pascale SÉBILLOT, Professeur des universités, INSA de Rennes	Examinatrice
Maximilien SERVAJEAN, Maître de Conférences, Université Paul-Valéry Montpellier 3	Examinateur
Sandra BRINGAY, Professeur des universités, Université Paul-Valéry Montpellier 3	Directrice
Caroline MOLLEVI, Chargée de recherche HDR, Université de Montpellier	Directrice

UNIVERSITÉ DE MONTPELLIER

janssen Horizon

Résumé

On retrouve de nombreux phénomènes de controverse dans les médias sociaux. Ils sont définis comme des sujets associés à un désaccord entre les utilisateurs et portent généralement sur un événement spécifique, impliquant une grande divergence d'opinions, des déclarations ne faisant pas l'unanimité, des débats, des contestations, jusqu'à des polémiques autour de sujets sensibles. Mieux comprendre les phénomènes de controverse dans les médias sociaux permet d'analyser les tendances sociales, l'évolution de l'opinion publique autour de divers sujets, de surveiller la propagation de la désinformation ou la réputation des personnalités publiques, etc. Dans ce contexte, la controverse dans les médias sociaux a été étudiée de manière intensive dans la littérature scientifique et récemment par des approches automatiques. L'objectif de cette thèse a été d'étudier la controverse en s'appuyant à la fois sur les propriétés textuelles des textes produits par les utilisateurs, mais également sur les propriétés structurelles de ces sujets issus des interactions des utilisateurs.

L'avènement de l'apprentissage automatique et plus spécifiquement de l'apprentissage profond a conduit à une amélioration de la représentation des données de tout type, à l'aide de techniques d'apprentissage plus complexes. Dans cette thèse, nous nous intéressons à deux types de domaines en particulier :

- Le traitement automatique de la langue (TAL), afin d'apprendre à représenter le texte à partir de modèles complexes, tels que BERT.
- Le traitement des graphes, permettant d'apprendre à représenter des données non structurées telles que les graphes. Nous étudions pour cela les progrès faits dans l'apprentissage des réseaux de neurones graphiques (GNN).

Dans cette thèse, nous présenterons nos travaux à travers quatre parties : (**1**) un état de l'art autour des méthodes d'analyse de la controverse ainsi que des réseaux de neurones graphiques, (**2**) une méthode expliquant les sujets controversés en analysant les textes produits par les communautés sur Twitter, (**3**) une méthode de prédiction des sujets controversés à partir des réponses entre utilisateurs sur Reddit, (**4**) une méthode basée sur les graphes et le texte pour quantifier la controverse des sujets sur Twitter, en se basant sur la polarisation des utilisateurs autour de communautés.

Mots-clés –Apprentissage automatique, Réseaux de neurones graphiques, Traitement automatique de la langue, Détection automatique, Controverse, Explicabilité, Médias sociaux

Abstract

Controversy is a common phenomenon in social media. They are defined as subjects associated with disagreement between users, and generally relate to a specific event, involving a wide divergence of opinions, statements not unanimously agreed upon, debates, challenges, even polemics around sensitive subjects. A better understanding of controversy in social media makes it possible to analyze social trends, the evolution of public opinion on various subjects, to monitor the spread of misinformation or the reputation of public figures, and so on. In this context, social media controversy has been extensively studied in the scientific literature and recently by automatic approaches. The aim of this thesis was to study controversy by drawing on both the textual properties of user-generated texts, and the structural properties of these topics derived from user interactions.

The advent of machine learning, and more specifically deep learning methods, has led to improved representation of data of all types, using more complex learning techniques. In this thesis, we focus on 2 types of domain in particular :

- Natural language processing (NLP), in order to learn complex textual representation using deep models based on attention mechanisms, such as BERT.
- Graph processing, in order to learn how to represent unstructured data such as graphs. To this end, we are studying the progress made in graph neural networks (GNN), and the methods derived from them.

In this thesis, we will present through four parts : (**1**) the state of the art in controversy analysis methods and graph neural networks, (**2**) a method to explain controversial topics by analyzing content of Twitter communities, (**3**) a method for early prediction of controversial posts from Reddit discussion between users. (**4**) a graph and text based method to quantify controversial topics on Twitter, based on user polarization around communities.

Keywords – Machine learning, Graph neural networks, Natural language processing, automatic detection, Controversy, Explainability, Social media

TABLE DES MATIÈRES

CHAPITRE 1

INTRODUCTION

Sommaire

Ce chapitre est une introduction générale à cette thèse. Il décrit le contexte dans lequel s'inscrit cette recherche et résume ensuite la problématique traitée et les contributions scientifiques mises en œuvre.

1.1 Contexte

1.1.1 La controverse

La popularité des réseaux sociaux (Twitter, Facebook, Instagram, Reddit, etc.) conjuguée à leur facilité d'utilisation, souvent sur téléphones mobiles (smartphones), ont fortement contribué à augmenter la connectivité des personnes. Une étude réalisée en 2017 [Wu+17] indique qu'il y a plus de deux milliards d'utilisateurs dans le monde. Ce nombre atteint les 6,8 milliards d'utilisateurs en 2023 [1]. Les sujets de discussion sont divers et peuvent concerner des thèmes comme la santé, le changement climatique, l'éthique, la culture, la religion, la politique, etc. Pour chaque sujet, de très nombreuses opinions sont exprimées, partagées, critiquées, etc. Parmi ces sujets, on trouve de très nombreuses controverses comme l'obligation de la vaccination contre la covid-19, le droit des individus à mettre fin à leur vie dans certaines circonstances difficiles, le port des armes ou encore le changement climatique.

La diversité, voire la contradiction, des opinions exprimées en quantité sur un sujet donné produit ce qu'il est convenu d'appeler un sujet controversé. Les débats autour des sujets controversés sont parfois houleux, voir violents et mettent en évidence des désaccords profonds entre des communautés. Différentes définitions de la controverse ont été proposées. Dans le Larousse [2], une controverse est définie comme "une discussion suivie sur une question, motivée par des opinions ou des interprétations divergentes". Ainsi, un sujet controversé suscite des opinions fortes et des désaccords qui montrent l'absence de consensus sur la question traitée. Une controverse est considérée dans certains cas comme tout contenu qui attire aussi bien des commentaires positifs que des commentaires négatifs [HL19]. Une controverse est définie aussi comme un phénomène de polarisation [Ras+21], c'est-à-dire un processus de concentration des individus autour d'opinions opposées. Il est important de souligner que la controverse peut être contextuelle [JDA17]. Un sujet peut être controversé dans une communauté, une région, un pays et ne pas l'être ailleurs. Un sujet peut devenir controversé alors qu'il ne l'était pas dans le passé. Il n'existe donc pas de définition complétement consensuelle de la controverse. Certaines se basent sur les sentiments exprimés, d'autres sur la polarisation des communautés, ou encore sur la prise de position des utilisateurs. La controverse révèle également des comportements utilisateurs spécifiques selon les médias sociaux, que ce soit au niveau des nombreuses interactions ou du contenu des messages. Ces deux aspects seront étudiés dans cette thèse.

Un exemple de discussion sur deux réseaux sociaux différents est illustré dans la figure 1.1. La partie supérieure de la figure concerne une discussion sur le réseau

1. https ://www.oberlo.com/statistics/how-many-people-have-smartphones
2. https ://www.larousse.fr/dictionnaires/francais/controverse/18941

Reddit sur le thème de la peine de mort. Un utilisateur Reddit peut réagir à un message initial ou à un commentaire sur un message pour le supporter, s'y opposer ou tout simplement exprimer sa neutralité. La discussion sur Reddit peut être représentée par un arbre de discussion ("post-comment tree" en anglais) dont la racine correspond au message initial et les autres nœuds correspondent aux différents commentaires. La partie inférieure de la figure 1.1 correspond à une discussion Twitter sur le discours du premier ministre Netanyahu du 3 mars 2015. Les utilisateurs retweetent des messages pour exprimer leur accord, et créent de nouveaux tweets pour exprimer leurs propres opinions. Il est important de souligner que les réseaux sociaux, en plus de différer dans les politiques de gestion de la vie privée et dans la qualité des informations fournies, différent aussi par les fonctionnalités qu'elles mettent à disposition des utilisateurs. Ces actions sont fondamentales dans l'analyse, la détection et la quantification de la controverse. Ainsi, l'action retweet est souvent considérée comme un moyen de supporter le contenu d'un tweet. Une telle action n'est pas possible dans Reddit où l'opinion d'une personne est tout simplement exprimée sous la forme d'un texte (commentaire ou post). La conception d'une méthode de détection de la controverse qui soit valable sur différents réseaux sociaux constitue un challenge. C'est la raison pour laquelle la plupart des méthodes de détection de la controverse se focalise sur un réseau spécifique.

La détection de sujets controversés offre plusieurs avantages. Elle permet entre autres de mettre en évidence la diversité de points de vue, poussant certaines personnes à sortir de leurs zones de confort pour examiner d'autres façons de considérer le sujet traité et revoir d'éventuels préjugés. La détection et la quantification de la controverse permet aussi de saisir la complexité du sujet et la difficulté de trouver ou pas des consensus. Cela est d'autant plus vrai lorsque se manifeste le phénomène de l'écho de chambre médiatique ("echo-chamber" en anglais). Il s'agit de l'amplification et du renforcement continu de contenus similaires par des systèmes de recommandation, pour augmenter la crédibilité de certaines informations.

L'étude de la controverse joue un rôle crucial dans la lutte contre la désinformation et la compréhension de la dynamique sociétale. Des travaux tels que ceux menés par Vosoughi, Roy et Aral [VRA18] ont démontré comment de fausses informations se propagent rapidement en exploitant les fractures de l'opinion publique. En analysant les débats controversés, les chercheurs peuvent identifier les discours trompeurs et les sources de désinformation, et mettre en place des stratégies pour atténuer leur diffusion. L'étude des controverses offre également des avantages aux entreprises cherchant à évaluer leur popularité auprès de leurs clients. Les travaux de Chen et al. [Che+19] soulignent comment l'analyse des débats en ligne peut fournir des informations précieuses sur la perception des marques et les attentes des consommateurs. En surveillant les discussions controversées liées à leurs produits ou services, les entreprises peuvent adapter leurs stratégies de communication et d'engagement pour maintenir une image positive et renforcer la confiance du public. L'étude de la controverse présente un intérêt tant pour la société en général, en l'aidant à lutter contre la désinformation, que pour les entreprises, en les guidant vers des décisions plus éclairées. Dans cette thèse, nous allons mettre en place des méthodes basées sur les avancées récentes de l'intelligence artificielle afin d'analyser la controverse sur les médias sociaux.

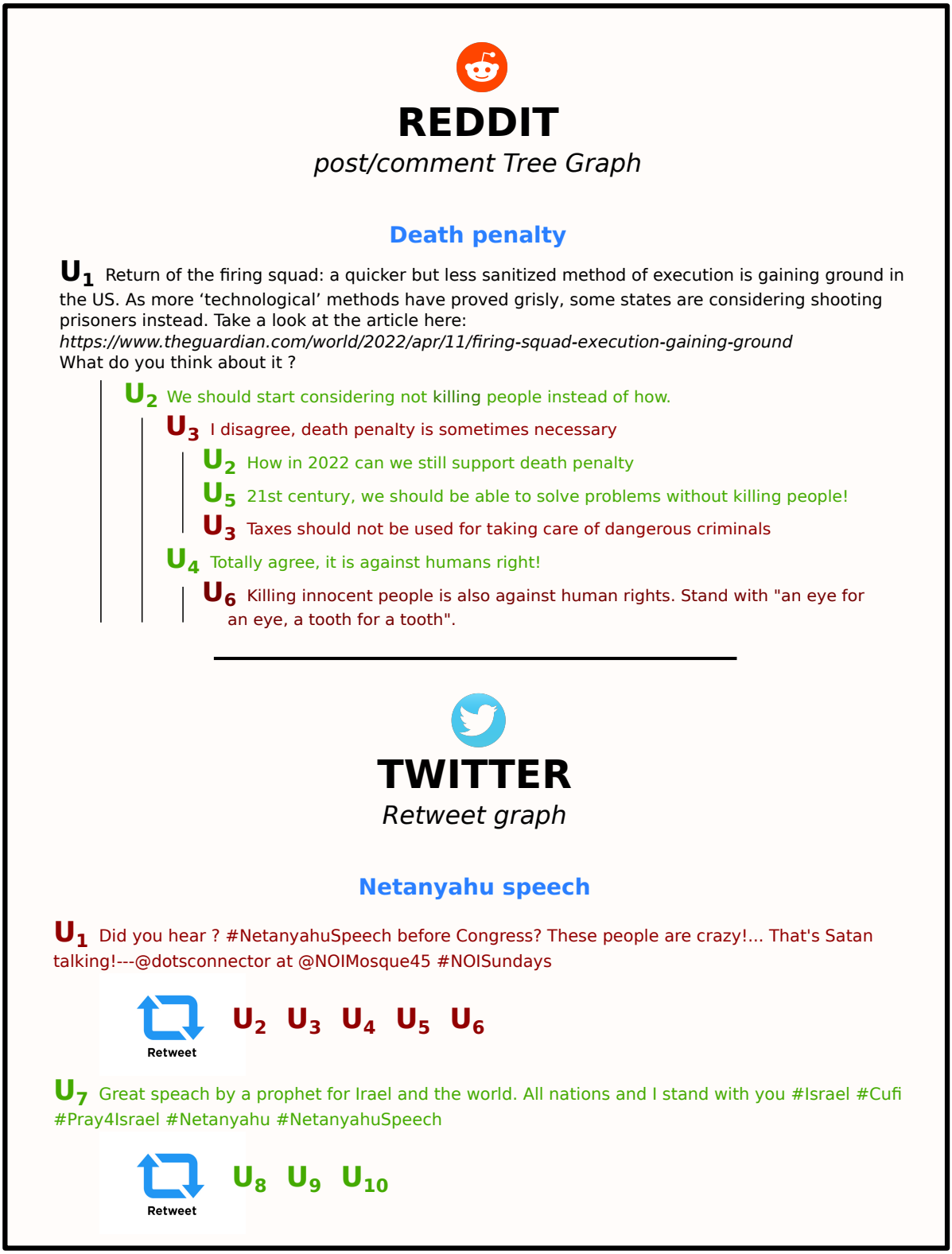

FIGURE 1.1 – *Exemples de discussions controversées sur Reddit et Twitter. Partie haute : une discussion sur Reddit où le post initial provoque un débat et les utilisateurs interagissent entre eux sous forme de commentaires. Partie basse : une discussion sur Twitter où les utilisateurs expriment une opinion commune via l'action de retweet. Ainsi, il y a moins de discussions sur Twitter.*

1.1.2 L'intelligence artificielle

Nous assistons ces dernières années à l'émergence forte et continue de l'intelligence artificielle (IA). Celle-ci envahit et envahira de plus en plus notre quotidien et connait incontestablement un essor impressionnant dans l'industrie, tous secteurs d'activités confondus. L'IA vise à effectuer des tâches simples ou complexes qui traditionnellement étaient réalisées par des êtres humains. Son impact sur notre société et ses diverses dimensions culturelles, sociétales ou économiques s'accroît au fil du temps. Les applications autour de l'IA incluent à titre d'exemple le domaine médical, la traduction automatique, la reconnaissance faciale, la génération de textes, la conversation, la voiture autonome, la détection de fraudes et d'anomalies, la détection de spams, etc.

Le concept de l'IA n'est pas nouveau et ses prémisses remontent aux années 50. Les systèmes experts appelés aussi systèmes à base de règles issus d'une approche symbolique logique ont connu un engouement dans les années 80. Ils permettaient de réaliser des prédictions par l'identification de règles de production (appelées aussi base de connaissances) qui établissaient des liens entre les faits et les résultats. Cette approche a très vite montré ses limites par la difficulté de disposer des règles dans des cas d'utilisation de plus en plus complexes. Une seconde approche appelée logique connexionniste a aussi vu le jour pratiquement dans les mêmes années 50. Elle s'appuyait sur le fonctionnement du cerveau humain, et l'apprentissage automatique ("Machine learning" [Lu+23] en anglais). Ce terme désigne les techniques visant à exploiter de grosses quantités de données pour y découvrir des motifs ou régularités et en tirer des prédictions en se basant sur des modèles statistiques. On distingue les méthodes supervisées et les méthodes non supervisées. Les méthodes supervisées nécessitent que les données d'entrainement soient labellisées. Elles permettent de réaliser des tâches de classification (comme prédire si un message est un spam ou non) et des tâches de régression (comme la prédiction d'occurrence d'un événement). Les méthodes non supervisées utilisent des données d'entrainement non labellisées. Elles permettent essentiellement de réaliser des tâches de regroupement ("clustering" en anglais). Les méthodes semi-supervisées correspondent à une combinaison des deux précédentes où seule une petite partie des données d'entrainement est labellisée. Les méthodes d'apprentissage par renforcement ne nécessitent pas de données d'entrainement et apprennent de leurs propres erreurs et essayent de maximiser leurs récompenses. L'approche connexionniste qui fut tout d'abord délaissée constitue désormais la base de la plupart des modèles d'IA d'aujourd'hui, notamment grâce aux progrès technologiques réalisés par l'informatique et plus particulièrement par l'augmentation de la puissance de calcul des machines. L'IA issue de cette approche connexionniste est diverse. Dans cette thèse, nous nous intéresserons à trois principaux aspects :

- Apprentissage profond ("Deep learning" en anglais) [HY18]. L'apprentissage profond est une technique d'apprentissage automatique avancée pour la réalisation de tâches complexes comme la reconnaissance de visage et la traduction entre langages. Cette approche, à la manière des neurones du cerveau humain, se base sur des réseaux de neurones artificiels contenant plusieurs couches de neurones.

- Les réseaux de neurones graphiques (“Graph Neural Networks” ou GNN en anglais [Wu+21]) sont des modèles d’apprentissage automatique basés sur des techniques d’apprentissage profond. Ils agissent sur des données de type graphes et offrent divers avantages dont principalement l’amélioration des performances sur des grands graphes et la réduction de la haute dimensionnalité des graphes. Les GNN représentent les nœuds d’un graphe sous la forme d’un vecteur de plongement dans un espace euclidien (“embedding” en anglais). Ces vecteurs de représentations ainsi obtenus permettent d’effectuer diverses tâches comme la classification des nœuds, la prédiction des liens entre nœuds, et la génération de graphes. Les GNN ont été utilisés dans différents domaines et incluent l’identification de protéines, les systèmes de recommandation et la prédiction de trafic dans une zone urbaine.

- Le traitement du langage naturel (TAL), aussi appelé “Natural Language Processing” (NLP) en anglais [Tre+23]), est une autre composante de l’IA se concentrant sur les méthodes et techniques d’analyse des textes. Les techniques d’apprentissage automatique ont beaucoup fait évoluer le traitement du langage. Les premières évolutions sont simples, et incluent des représentations telles que l’encodage “one-hot” des mots ou la mesure de la cooccurrence TF-IDF. Le plongement lexical, ou représentation lexicale, des mots (“Word embedding” en anglais) a constitué une avancé importante. Son principe est de représenter chaque mot du vocabulaire considéré dans un espace vectoriel de taille fixe. Il réduit la dimension et est capable de capturer les relations de similarité des mots. Il est réalisé par un réseau de neurones profond non supervisé appelé auto-encodeur. On distingue les plongements lexicaux simples pour une représentation statique et non contextuelle du sens des mots (Word2vec [Mik+13], GloVe [PSM14], FasText [Mou22]), et le plongement lexical contextuel qui représente le sens du mot en fonction de son contexte d’utilisation. Le TAL a connu un grand essor avec l’avènement des architectures à base de Transformers et du mécanisme d’attention. Ces derniers combinent les techniques de plongements lexicaux avec le mécanisme d’attention dans le processus d’apprentissage d’encodage de la sémantique du langage. Le modèle BERT [Dev+19] proposé par Google et le modèle GPT [Bro+20] par openAI sont les modèles de référence. Ces méthodes se séparent en deux phases d’entrainement, une phase de pré-entrainement des représentations du langage, et une phase d’apprentissage par transfert (“Transfer learning” en anglais) et ont été utilisées pour différentes tâches de TAL.

1.1.3 Objectif de la thèse

Les données textuelles constituent une part très importante des données disponibles dans les réseaux sociaux. Elles incluent donc l’expression textuelle dans une langue ou une autre des opinions des utilisateurs relatives à un sujet. En plus d’être faciles à collecter et à stocker contrairement aux vidéos et aux images, les données textuelles sont présentes en grande quantité. Cependant, sur un réseau social tel que Twitter, le contenu des messages ne correspond qu’à une partie des informations disponibles sur le sujet. Par exemple, sur le sujet controversé autour du discours au congrès de Nancy Pelosi à propos de la mise en accusation (“impeachment” en anglais) de Donald

Trump en décembre 2019, et à partir des données (tweets et retweets) récoltées sur Twitter par Zarate et al. [Zar+20], moins de 10% des utilisateurs participant au sujet diffusent du contenu. Les interactions entre utilisateurs, et donc les informations structurelles de la discussion, sont des informations pertinentes. Nous considérons que la combinaison des contenus textuels avec les données structurées issues des interactions entre utilisateurs, peut améliorer l'analyse de la controverse. Ainsi, l'objectif de cette thèse est d'exploiter les dernières avancées de l'IA et plus particulièrement des réseaux de neurones, des réseaux de neurones graphiques (GNN) et des techniques de NLP pour les besoins d'explication, de détection et de quantification de la controverse dans les médias sociaux. Nous exploiterons la dimension textuelle des opinions des utilisateurs que nous combinerons avec la dimension interaction entre utilisateurs pour ces trois besoins.

1.2 Problématiques de recherche et contributions

1.2.1 Problématiques de recherche

Les contenus controversés suscitent des opinions et des interrogations différentes des utilisateurs et impliquent des interactions d'intensité variable. Expliquer, détecter et quantifier la controverse sont des tâches difficiles et constituent de réels challenges. Les informations structurelles extraites à partir des interactions entre utilisateurs ont été largement utilisées avec les techniques classiques de partitionnement de graphes pour les besoins de détection de la controverse. Certaines approches, même si elles ne sont pas nombreuses, exploitent aussi le texte du contenu des discussions. Les avancées récentes, notamment autour des réseaux de neurones graphiques (GNN) et des modèles de langages larges, peuvent être combinées pour expliquer, détecter et quantifier la controverse.

Dans ce travail de thèse, nous adoptons les techniques de GNN et de modèles de langages larges, et étudions les questions de recherche suivantes :

- **Q1** : Comment **expliquer** la controverse du point de vue des communautés utilisateurs ? Quelles sont les caractéristiques du texte qui peuvent être qualifiées de prédictives des communautés controversées ?
- **Q2**. Comment **détecter** automatiquement la controverse autour de discussion entre utilisateurs, en utilisant les avancées récentes en matière de GNN et de modèles de langages ? Il s'agit d'abord d'identifier dans quelle mesure les GNN arrivent à détecter la controverse en utilisant uniquement l'information structurelle. Il s'agit ensuite de mesurer l'apport du texte dans cette détection.
- **Q3**. Comment **quantifier** la controverse sur les réseaux sociaux à partir de l'apport des utilisateurs à un sujet ? Nous quantifions la controverse, à partir de l'espérance de tirer un utilisateur aléatoirement participant à un sujet controversé. Ensuite, nous estimons ce score d'une manière empirique, à partir d'un modèle basé sur les GNN. Comme pour la détection, nous nous intéresserons au contenu de ces messages et aux relations entre utilisateurs.

1.2.2 Contributions

Dans chacune des contributions suivantes, nous nous intéresserons à la prise en compte du texte en plus de la structure des graphes d'interactions pour les trois tâches d'explication, de détection et de quantification. La combinaison des informations structurelles et textuelles, ainsi que l'application des GNN, sont nos fils conducteurs tout au long de cette thèse. Nous énumérons dans cette section les différentes contributions réalisées.

1. **Contribution C1 - Explication de la controverse par analyse des tweets des communautés**. En vue d'analyser quelles sont les caractéristiques textuelles les plus significatives contribuant le mieux à comprendre les communautés des sujets controversés, nous proposons une approche basée sur les deux considérations suivantes :

   - Exploration de la controverse du point de vue des communautés : nous considérons pour un sujet controversé deux communautés opposées dans leurs opinions sur le sujet. Puis, nous explorons la controverse sous l'angle de la classification d'un tweet dans une des deux communautés opposées en utilisant uniquement le contenu du tweet. Nous basons nos travaux sur un modèle de langage de type BERT. En plus d'un score de polarisation [Gar+18a], un score de controverse est calculé à partir du texte. Ce score s'appuie sur les performances du modèle de langage à pouvoir correctement classifier les tweets d'un jeu de validation. La précision du modèle BERT, acc_bert est utilisée comme score quantifiant la controverse du point de vue du texte. Un score acc_bert élevé correspond à la capacité du modèle à correctement prédire les communautés auxquelles appartiennent les tweets. L'étiquetage des tweets dans la phase d'apprentissage est basé sur l'algorithme metis [KK95] de partitionnement des utilisateurs présents dans le graphe de Retweet.
   - Utilisation de la méthode SHAP : Nous considérons la classification de tweets comme un jeu collaboratif, où les différentes caractéristiques du texte jouent le rôle des joueurs. La méthode SHAP offre alors un cadre pour mesurer équitablement les contributions des différentes caractéristiques du texte à la prédiction du classifieur. Différents ensembles de caractéristiques textuelles issues de TF-IDF et de BERT et de caractéristiques conceptuelles issues de la ressource LIWC [Boy+22] sont explorées.

   Les résultats expérimentaux menés sur des sujets controversés et non controversés montrent que le texte possède des caractéristiques intéressantes et complémentaires aux interactions entre utilisateurs pour l'analyse de la controverse. Sur les sujets controversés, l'analyse des tokens (mots) ayant une contribution importante sur la classification des tweets par le modèle BERT montre que celui-ci capture bien les caractéristiques textuelles liées aux communautés. Les résultats expérimentaux montrent que les caractéristiques conceptuelles issues de LIWC peuvent aussi aider à caractériser des tendances autour d'une communauté. À notre connaissance, nos travaux constituent la première initiative d'analyse des contributions sur la classification des tweets dans les communautés controversées

en se basant uniquement sur le texte. Notre approche ouvre aussi de nouvelles perspectives, comme la modélisation des utilisateurs en fonction des textes qu'il diffuse. Notre travail constitue aussi la première initiative utilisant la méthode SHAP pour analyser l'impact des caractéristiques du texte dans le domaine de la controverse.

2. **Contribution C2 - Détection des discussions controversées par l'utilisation des réseaux de neurones graphiques appliqués aux interactions utilisateurs et aux contenus**. Nous proposons une approche de détection automatique de la controverse dans les médias sociaux basée sur les GNN. Notre approche modélise une discussion entre utilisateurs sous la forme d'un graphe d'utilisateurs. Celui-ci représente les interactions entre utilisateurs, augmentées d'une représentation du contenu textuel des messages des utilisateurs extrait à partir de modèles de langages de type BERT ou LSTM. Nous considérons la détection de la controverse comme un problème de classification de graphes (le graphe est soit controversé, soit non controversé). Nous séparons ces travaux en deux parties :

   - Détection des discussions controversées à l'aide des GNN : nous proposons deux stratégies de détection, basées sur les GNN : (1) La stratégie dénommée HRL-GCN ("Hierarchical Representation Learning based on GCN" en anglais), exploite la structure hiérarchique pouvant exister dans un graphe d'utilisateurs et encode la totalité du graphe grâce à des techniques de GNN et de manière itérative et hiérarchique. Pour ce faire, nous exploitons l'approche DIFFPOOL [Yin+18] de classification des graphes combinée aux GCN [KW17]. (2) La stratégie dénommée ARL-GAT ("Attention Representation Learning based on GAT" en anglais), exploite le mécanisme d'attention dans le processus d'encodage du graphe par les GNN. Un tel mécanisme permet à chaque nœud utilisateur du graphe de juger quels sont les nœuds voisins les plus importants dans l'agrégation des informations. L'approche proposée a été complètement implémentée et des évaluations expérimentales sont réalisées en considérant le forum de discussion Reddit.

   - Généralisation de l'approche sur différents types de données : des expérimentations sont effectuées sur le réseau social Twitter, sur des graphes d'interactions différents (graphe de Retweet). Notre approche fonctionne sur d'autres types de réseaux sociaux dès lors qu'il est possible de générer un graphe d'utilisateurs.

   Les expérimentations se focalisant sur la détection de la controverse sur Reddit sont prometteuses. La stratégie HRL-GCN avec une couche de regroupement montre des performances supérieures, surpassant même l'état de l'art dans certains jeux de données. L'intégration d'informations textuelles issues de modèles tels qu'un BERT combiné à un bi-LSTM améliore la précision globale sur la majorité des jeux de données, soulignant leur pertinence pour la détection de la controverse. Cependant, l'inclusion de caractéristiques liées aux sentiments exprimés se révèle moins efficace, suggérant que d'autres éléments sont nécessaires pour une caractérisation complète de la controverse. Ces résultats mettent en évidence l'apport des informations textuelles pour une détection de la contro-

verse dans les discussions en ligne. Dans le cas du réseau Twitter, l'étude s'est concentrée sur l'analyse des graphes de retweets de 15 sujets, distinguant neuf sujets controversés et six non controversés. Trois techniques d'échantillonnage ont été utilisées pour gérer la complexité des graphes, et nous avons utilisé la méthode de validation croisée "leave-one-out". La méthode HRL-GCN associée à des caractéristiques textuelles extraites via BERT a obtenu les meilleurs résultats, montrant sa capacité à généraliser malgré des données en nombre limité. Cependant, l'efficacité a été moindre sur certaines méthodes d'échantillonnages, en raison du nombre restreint de graphes et de la grande taille des nœuds utilisateurs.
Ainsi, les expérimentations réalisées renforcent l'idée que les interactions entre utilisateurs constituent une information pertinente dans la détection de la controverse. Elle met aussi en évidence la complémentarité du texte et des interactions dans le cas de la plateforme Reddit en améliorant la précision de la détection. À notre connaissance, nos travaux constituent la première initiative exploitant en profondeur l'utilisation des GNN pour les besoins de la détection de la controverse. Seule l'approche proposée par Zhong et al. [Zho+20] utilise des GCN ("Graph Convolutional Network" en anglais) dans le contexte de la controverse sur Reddit. Notre approche diffère fondamentalement de la leur par le fait que nous accordons de l'importance à l'utilisateur ainsi qu'aux textes qu'il écrit alors que leur approche modélise le texte indépendamment de l'auteur.

3. **Contribution C3 - Quantification de la controverse sur Twitter d'un point de vue utilisateur.** Nous proposons dans cette thèse une nouvelle approche, utilisant les GNN, basée sur la probabilité d'un utilisateur de participer à un sujet controversé, afin de quantifier la controverse sur Twitter. Notre approche modélise les interactions entre les utilisateurs sous la forme d'un graphe de retweets, liant un utilisateur partageant les mêmes idées que l'auteur d'un tweet. Les nœuds utilisateurs sont initialisés à partir d'une représentation agrégée du contenu de leurs tweets originaux, extrait à partir d'un modèle de langage de type BERT, intégré dans le modèle. Notre approche se divise en deux principes :

   - Quantification de la controverse : nous proposons une modélisation de la quantification de la controverse pour un sujet Twitter, à partir de l'espérance d'un utilisateur tiré aléatoirement de participer à un sujet controversé. Afin d'estimer ce score, nous présentons une approche empirique, en se basant sur l'estimation des probabilités pour un ensemble d'utilisateurs de participer à un sujet controversé. Nous montrons aussi que minimiser l'erreur de cet estimateur avec un certain type de fonction de perte revient à maximiser notre score de quantification.

   - Prédiction de la probabilité d'un utilisateur de participer à un sujet controversé : pour estimer cette probabilité, nous utilisons à la fois les informations textuelles des utilisateurs (leurs tweets) et les informations structurelles (voisinage du graphe). Comme pour la contribution (C2), un modèle de type GNN est utilisé afin d'apprendre les nouvelles représentations utilisateurs avec ces deux types d'information. Cependant, contrairement à la contribution (C2), l'apprentissage de la représentation textuelle des utilisateurs est

inclus dans l'apprentissage de ce modèle afin de l'affiner. Nous avons testé deux modèles : (1) une méthode de convolution $GraphSAGE$ [HYL17], et (2) une méthode de convolution GAT [Vel+18] basée sur le mécanisme d'attention. Afin de comparer les performances de nos différentes méthodes, nous établissons un protocole d'évaluation, se basant sur un jeu de test comprenant des sous-graphes centrés sur chaque utilisateur avec différents niveaux k de voisinage. Nous étudions aussi l'évolution de ce score à différents k. Enfin, afin de vérifier la qualité de notre score, nous comparons aussi la qualité de la séparation des scores (AUC-ROC) avec les approches de la littérature.

Les travaux de ce chapitre étant encore en cours de réalisation, nous présentons des résultats préliminaires en testant notre approche uniquement sur des graphes entiers, pour la tâche de prédiction des probabilités d'un utilisateur à participer à un sujet controversé. Notre modèle basé sur le modèle GAT avec deux couches et utilisant l'agrégateur $MEAN$, obtient alors les meilleures performances. Ensuite, des expérimentations sont effectuées afin d'évaluer la qualité de notre approche pour la quantification de la controverse à partir de 10 sujets représentant des sujets controversés et non controversés équitablement. Nous comparons notre score de quantification à deux scores de la littérature [Gar+18a], ainsi qu'à des variantes de notre score, n'utilisant que les informations textuelles (les tweets) ou structurelles (le degré des nœuds est utilisé en caractéristiques d'entrée) des données. Ces expériences montrent la qualité de notre score pour correctement séparer les sujets controversés et non controversés.

Notre approche ouvre aussi de nouvelles perspectives au niveau de la quantification de la controverse. Elle est indépendante du nombre d'utilisateurs utilisés dans le calcul. Des études autour de l'impact de certains utilisateurs ou de prévisions de l'évolution de la controverse dans le futur constituent des perspectives intéressantes.

1.3 Organisation de la thèse

Le reste de ce manuscrit de thèse est constitué de la manière suivante :

- Le chapitre 2 met en avant les travaux effectués autour de l'analyse, la quantification et la détection de la controverse sur le web, et notamment dans les réseaux sociaux. Nous présentons également dans ce chapitre les recherches autour de l'application de l'apprentissage profond sur les données non structurées que sont les graphes [Wu+21] avec les réseaux de neurones graphiques (GNN).
- Le chapitre 3 concerne les travaux sur l'explication des communautés controversées sur Twitter. Nous présentons tout d'abord une méthode permettant de quantifier la controverse de manière plus efficace en s'appuyant sur la capacité d'un modèle à prédire les bonnes communautés des tweets. Nous présentons ensuite une méthode pour l'explication de la contribution des caractéristiques textuelles (et conceptuelles) à la prédiction de ces communautés par nos différents modèles, en se basant sur la méthode SHAP.

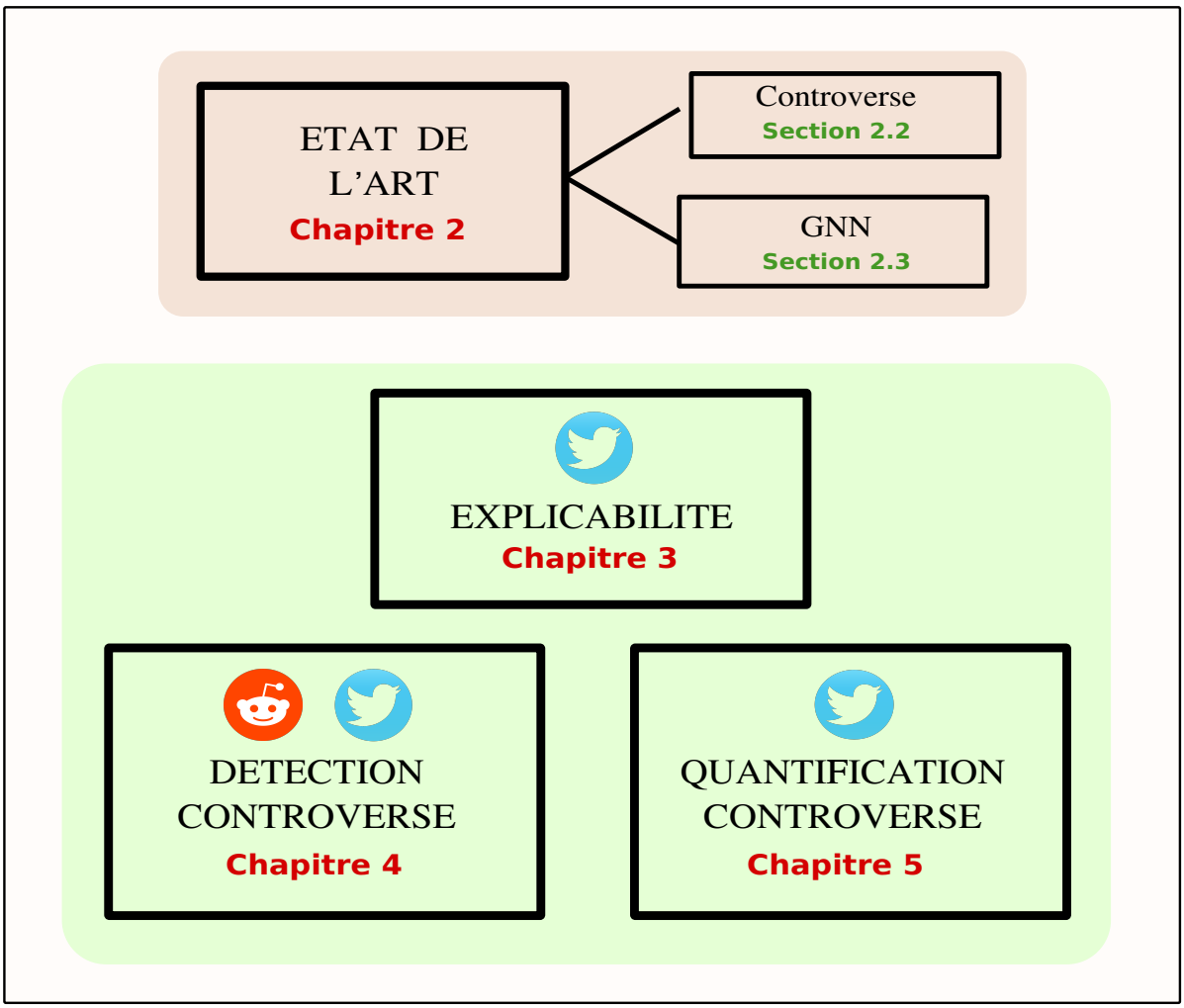

FIGURE 1.2 – *Vue d'ensemble des différents chapitres abordés dans ce mémoire. En orange, la partie concernant l'état de l'art. En vert, on retrouve les travaux effectués durant cette thèse abordant l'analyse de la controverse, avec les médias sociaux respectifs étudiés.*

- Le chapitre 4 se concentre sur la détection des posts (ou discussions) controversés sur Reddit. Nous présentons un modèle se basant sur des GNN et prenant en compte les messages des utilisateurs. Nous représentons notre problème comme une tâche de classification de graphes. Nous comparons deux approches pour représenter le graphe sous la forme d'un vecteur, dans un espace euclidien, afin d'appliquer ensuite la classification des posts. Enfin, nous présentons aussi des expérimentations sur des graphes utilisateurs différents, provenant de Twitter (graphes de retweets), afin d'étudier les capacités de notre modèle à généraliser sur tout type de données.
- Le chapitre 5 se concentre sur la quantification de la controverse d'un point de vue utilisateur. Nous nous appuyons sur une solution théorique pour quantifier la controverse, puis sur une approche empirique afin d'estimer ce score à partir de la probabilité estimée des utilisateurs à participer à un sujet controversé. Pour cela, nous présentons un modèle, basé aussi sur les GNN et la représentation textuelle des utilisateurs (à l'aide de BERT). Nous montrons que la combinaison des informations structurelles et textuelles appliquée à notre modèle donne de meilleurs résultats que l'utilisation des informations structurelles et textuelles séparément.
- Le chapitre 6 conclut cette thèse en résumant les contributions produites, et en présentant les perspectives de nos travaux. Ces dernières concernent les extensions possibles de nos travaux sur la quantification de la controverse, ainsi que sur la compréhension et l'explication de la controverse, à partir de l'utilisation de nouvelles approches autour des GNN.

La figure 1.2 regroupe les principaux chapitres, ainsi que le plan adopté afin de présenter le mémoire de cette thèse sur l'analyse de la controverse sur les médias sociaux.

Remerciement

Ce travail a été financé par des subventions provenant du fonds de dotation Janssen Horizon. Le projet s'étend d'octobre 2020 à octobre 2024, et le coordinateur scientifique entre le laboratoire et Janssen Horizon est Maximilien Servajean. Lors de cette thèse, l'accès aux ressources HPC de l'IDRIS a été accordé dans le cadre de l'allocation *AD011012604* faite par GENCI.

CHAPITRE 2

ÉTAT DE L'ART

Sommaire

2.1 Introduction

Dans ce chapitre, nous abordons les travaux de la littérature en lien avec l'étude de la controverse en ligne, de la détection et de la quantification des sujets controversés dans la section 2.2. Ces tâches feront l'objet de travaux et d'études approfondies dans les chapitres 3, 4 et 5. Nous souhaitons comprendre quelles caractéristiques sont utiles pour l'analyse de la controverse pour ces différentes tâches, qu'elles soient de nature textuelles ou structurelles. L'accent est mis sur l'analyse des médias en ligne et des réseaux sociaux, en s'appuyant notamment sur l'apport des interactions des utilisateurs.

Ensuite, nous présentons les concepts associés aux représentations des données sous forme de graphes dans la section 2.3, ainsi que les concepts associés aux réseaux de neurones graphiques, de l'anglais Graph Neural Networks (GNN), afin de représenter les nœuds et le graphe, pour différentes tâches de classification. Nous souhaitons définir les GNN au sens large, ainsi que différentes tâches d'apprentissage (supervisés ou non) et cas d'applications. Finalement, nous nous focalisons sur les réseaux de neurones à convolution graphique, principalement utilisés lors des travaux présentés dans les chapitres 4 et 5.

Enfin, nous présentons dans la section 2.4 différents jeux de données utilisés dans la littérature. Une présentation approfondie est faite en particulier d'un jeu de données provenant du réseau social Reddit, utilisé dans nos travaux dans le chapitre 4, ainsi que de deux jeux de données provenant du réseau social Twitter, utilisés dans nos travaux dans les chapitres 3, 4 et 5.

Il est important de noter que l'ensemble des figures sont des productions originales qui uniformisent la manière de présenter les réseaux de neurones graphiques. Nous avons également produit un tableau récapitulatif rassemblant les principaux travaux réalisés autour de la controverse, décrits selon les tâches qu'ils permettent de réaliser et les méthodes utilisées.

2.2 Quantification, détection et explication de la controverse

La notion de controverse, définie dans la section 1.1.1, peut être représentée de différentes manières. Elle peut être associée à un sujet, une personne, un événement, un article, etc. Cette section s'attardera sur deux éléments prédominants du web d'aujourd'hui : les **pages web** (articles), écrites par un nombre limité d'utilisateurs, mais prenant en compte l'avis général -estimé par ces derniers- une population concernée par le sujet, et les **médias sociaux**, où tout utilisateur peut accéder et participer à différents sujets, controversés. La figure 2.1 présente différents types de contenu présents en ligne, sur les différentes plateformes étudiées dans cet état de l'art.

2.2.1 Analyse du contenu des pages web

Nous nous intéressons dans cette section aux approches qui analysent la controverse présente dans des articles publiés en ligne, sur différentes plateformes comme Wikipédia (voir section 2.2.1.1) ou médias de type journalistique (voir section 2.2.1.2).

2.2.1.1 Détection des pages Wikipédia controversées

Plusieurs travaux étudient la controverse sur le web, grâce à des sources telles que Wikipédia [1]. Wikipédia est une encyclopédie en ligne gratuite et collaborative qui permet à des utilisateurs du monde entier de créer et de modifier des articles sur une grande variété de sujets. Les pages (ou articles) provenant de Wikipédia peuvent être automatiquement étiquetées comme controversées ou non, en utilisant les "edit-wars" [2] et/ou les relations/citations entre les pages. Wikipédia présente l'avantage de définir un sujet en reflétant la vision de ces rédacteurs, et les sujets controversés y sont très présents [3]. Wikipédia est souvent utilisé pour étudier la controverse, car cette ressource contient beaucoup de données exploitables, incluant du texte et des métadonnées. Les métadonnées correspondent à toutes caractéristiques concernant la publication de l'article en question, tels que le nombre de révisions et d'éditeurs, le nombre moyen de révisions et d'éditeurs anonymes, de tags controversés, la moyenne d'éditions par éditeurs, le nombre d'éditions de type "retour en arrière", la taille de l'article, le nombre de liens vers d'autres articles, etc.

Kittur et al. [Kit+07] présentent une méthode basée sur l'apprentissage automatique pour prédire les conflits à partir du contenu des articles. Chacun des articles est représenté par le nombre d'éditions mineures, le nombre de révisions, la longueur de l'article, le nombre d'éditions par l'administrateur et le nombre d'éditions anonymes. Une mesure basée sur le nombre de révisions controversées (CRC) est introduite afin de labelliser les pages. Elle utilise l'historique des modifications pour mesurer le nombre total de révisions au cours desquelles l'étiquette "controversé" (fournie par le système de catégorisation de Wikipédia) a été appliquée à l'article. La CRC est calculée pour toutes les révisions de chaque article de Wikipédia. Une CRC supérieur à zéro signifie que l'article a fait l'objet d'une révision controversée. Ensuite, l'algorithme des machines à vecteurs de support (SVM) est utilisé pour prédire les scores CRC des articles à partir de ces caractéristiques. Un modèle de graphe d'inversion basé sur la relation d'inversion qui pourrait exister entre les utilisateurs sur un article particulier est également défini pour identifier les conflits. Le regroupement de ces graphes permet d'identifier les groupes d'utilisateurs et leurs opinions respectives.

Vuong et al. [Vuo+08] basent leurs travaux sur la connexion entre les utilisateurs en litige pour identifier automatiquement les articles controversés. La relation de "litige" (ou dispute), qui représente un désaccord entre les utilisateurs, est extraite de l'historique des modifications de l'article. Elle est quantifiée par le nombre de mots

1. www.wikipedia.org
2. Situation où plusieurs rédacteurs différents échangent des opinions divergentes autour d'un sujet (une page) sur Wikipédia.
3. *https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues*. Liste non exhaustive de sujets controversés sur Wikipédia.

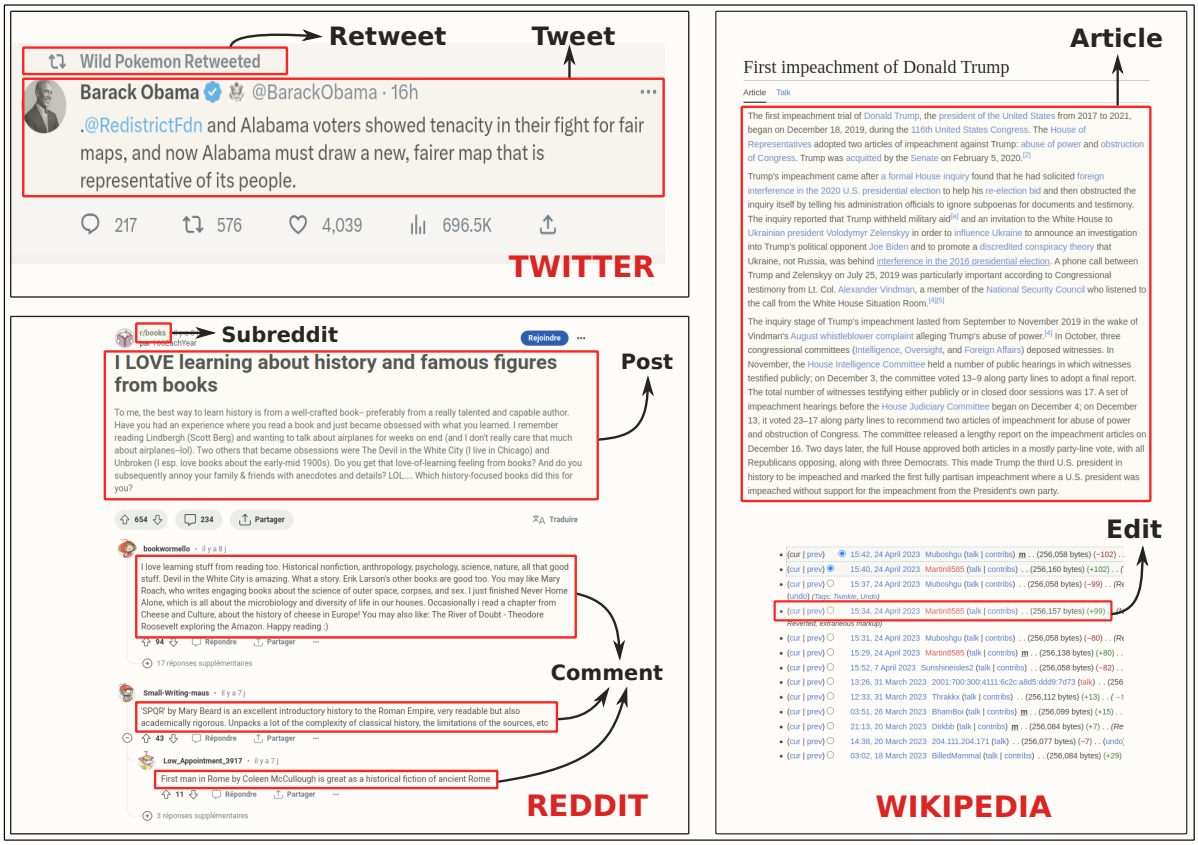

Figure 2.1 – *Exemple de contenus présents en ligne. Trois différentes plateformes sont présentées : Deux réseaux sociaux, Twitter et Reddit, ainsi qu'une encyclopédie en ligne, Wikipédia. Pour chaque plateforme, les différents types de contenus étudiés sont présentés.*

que les deux utilisateurs ont supprimé l'un de l'autre. Les articles Wikipédia sont alors représentés sous la forme d'un graphe bipartite où les nœuds correspondent à des paires ordonnées d'utilisateurs ou d'articles, et les arêtes dirigées relient les paires d'utilisateurs aux articles. Chaque arête met en évidence un désaccord entre deux utilisateurs sur un article donné, et est pondérée par le nombre de mots supprimés par un contributeur sur le texte d'un autre contributeur. Les contributions des utilisateurs aux articles sont également représentées par un second graphe bipartite dans lequel les nœuds correspondent soit aux utilisateurs, soit aux articles, et les arêtes dirigées relient les utilisateurs aux articles auxquels ils ont contribué. Ces arêtes sont pondérées par un poids correspondant au nombre de mots ajoutés par un utilisateur à l'article concerné. Deux modèles sont proposés pour classer les articles controversés. Le premier modèle considère la controverse d'un article du point de vue de son contributeur. Plus un article contient de tags de controverse, plus il est controversé. La controverse d'un article donné est définie comme le rapport entre le nombre total de litiges (calculé à partir du premier graphe bipartite) et le nombre total de contributions sur l'article de l'ensemble des utilisateurs (calculé à partir du second graphe bipartite). Le second modèle se concentre sur les litiges du point de vue de la nature de l'article. Deux hypothèses sont alors retenues : "Un article est plus controversé s'il contient plus de litiges entre des contributeurs moins controversés", et "Un contributeur est plus controversé s'il est engagé dans plus de litiges dans des articles moins controversés". Ce deuxième modèle est également proposé avec la possibilité de tenir compte de l'âge de l'article afin de réduire l'impact des litiges dans les nouveaux articles. L'âge d'un article est défini comme le nombre de révisions qu'il a subies.

Yasseri et al. [Yas+12] supposent dans leurs travaux que la détection des conflits sur Wikipédia devrait être multilingue et indépendante de la culture des participants. Ils proposent ainsi une méthode de détection des conflits basée sur les caractéristiques statistiques des éditions. Les changements réalisés par des éditeurs entre deux révisions sont identifiées en calculant et en comparant les différentes révisions à l'aide de la fonction de hachage cryptographique MD5. La mesure de controverse est ensuite calculée en utilisant des paires de changements sur un même article, puis en comparant le conflit entre toutes les paires de ce même article.

Ces méthodes se basent donc sur une classification binaire de la controverse. Jang et al. [Jan+16] proposent une autre perspective en étendant ces travaux à un modèle probabiliste, se basant toujours sur un modèle linguistique. Ce modèle est utilisé pour le classement des pages web, se rapprochant ainsi d'une tâche de quantification de la controverse. Les pages Wikipédia sont caractérisées par 3 scores WCF (de l'anglais "Wikipedia Controversy Features"), pré-calculés à partir des métadonnées, basées sur les analyses faites par Dori-Hacohen et Allan [DA15] dans leurs premiers travaux :

- **M**, basé sur les caractéristiques des révisions, servant à prédire la férocité d'une "edit-war".
- **C**, basé sur un modèle de régression entrainé sur les étiquettes controversées ou non de la page.
- **D**, représente la présence de l'étiquette "Litige" ou non.

À partir de ces scores, un score de controverse est calculé. Jang et Allan [JA16] améliorent ensuite cette approche. Pour cela, ils génèrent des requêtes pour découvrir les plus proches voisins d'une page. Puis, ils génèrent plusieurs requêtes à partir de paragraphes plus petits, mais plus cohérents du point de vue du document. Ensuite, ils modifient le score **M** des caractéristiques WCF pour parer l'éventualité qu'un document n'ait pas d'historique de révision, ce qui biaiserait ce score.

Dori-Hacohen, Jensen et Allan [DJA16] utilisent cette intuition dans leur approche, tout en exploitant les dépendances entre les articles pour détecter la controverse pour un article donné. Leur modèle s'inspire de l'inférence collective, supposant que la controverse se produit dans des voisinages de sujets apparentés. Leur approche suppose aussi qu'incorporer la similarité textuelle entre articles présente un vrai pouvoir prédictif, plus important que les hyperliens classiques, et promeut l'utilisation d'un réseau construit sur la base de la similarité du texte. Pour un article Wikipédia donné, le voisinage est calculé non pas à partir des liens, mais en utilisant une mesure de similarité entre l'article et les articles cités ou citant cet article. Ces documents sont décrits par leur contenu, en représentant le texte par un vecteur de fréquence TF-IDF de chaque terme utilisé. Le modèle sélectionne ensuite les top-k pages les plus proches, agrège les métadonnées Wikipédia servant de caractéristiques à chaque document en un vecteur, puis les classifient, à l'aide d'un algorithme de forêts aléatoires ("Random Forest" en anglais). L'approche combine donc l'utilisation du contenu textuel aux métadonnées des pages Wikipédia.

Les métadonnées, très utilisées auparavant, présentent néanmoins la controverse seulement à partir du comportement des contributeurs, et non d'un point de vue du contenu. Pour cela, plusieurs travaux ont analysé la controverse à partir de caractéristiques textuelles et se concentrent sur la sémantique du langage, en supposant que le contenu du texte puisse être utilisé comme outil de détection d'un sujet ou d'un message controversé.

Sznajder et al. [Szn+19] présentent une approche permettant de mesurer le degré de controverse d'un concept sur les pages de Wikipédia. Au lieu de s'appuyer sur les métadonnées de Wikipédia, les auteurs soutiennent que le contexte textuel immédiat d'un concept est révélateur de la controverse. Les concepts sont labellisés, controversés selon les tags associés. Ils sont ensuite représentés à l'aide de vecteurs de représentation de mots pré-entraînés (en anglais "word embedding") et définissent trois estimateurs de controverse basés sur différents algorithmes d'apprentissage automatique : la méthode des plus proches voisins, un modèle probabiliste (Naïve Bayes) et un réseau de neurones récurrents. Pour la méthode des plus proches voisins, le score final correspond à la proportion de concepts controversés dans le voisinage proche.

Wang et Cardie [WC16] proposent une approche d'analyse des sentiments au niveau de la phrase pour le problème de la détection des litiges. L'approche s'attache à représenter les caractéristiques lexicales et thématiques de la discussion. Les séquences de sentiments au niveau des phrases (très négatifs, négatifs, neutres, positifs, très positifs) exprimés dans les discussions sont identifiées et utilisées dans un classifieur binaire. Le classifieur a pour but de prédire si l'ensemble de la discussion est contesté. L'approche proposée a été évaluée sur un corpus regroupant les pages de discussion

des articles controversés et non controversés, comprenant notamment les discussions conflictuelles sur Wikipédia. Les différentes méthodes évoquées exploitent donc essentiellement les caractéristiques textuelles ainsi que les métadonnées sur les révisions de ces articles.

2.2.1.2 Détection des sujets controversés provenant des médias en ligne

Les médias en ligne, nécessitant une certaine objectivité (à la différence de Wikipédia), regroupent des articles mis en ligne par des journalistes et sont donc non modifiables par d'autres contributeurs une fois mis en ligne. Cependant, les internautes peuvent réagir à ces articles, se rapprochant peu à peu du fonctionnement d'un média social. De nombreuses approches se focalisent sur la recherche de désaccord entre utilisateurs [Yas+12 ; Kit+07 ; WC16].

Dans les médias en ligne, Sriteja, Pandey et Pudi [SPP17] s'intéressent à la détection d'articles controversés, en se focalisant à la fois sur le contenu textuel de l'article en question, et sur l'analyse de sentiments des réactions en rapport à l'article sur les réseaux sociaux des internautes. Les auteurs supposent que la présence de différentes émotions et l'intensité (représentée par une forte concentration de commentaires et d'utilisateurs impliqués) d'une discussion sont corrélées à la controverse. Se basant sur des articles de différents médias en ligne et de Facebook pour les réactions, ils proposent une méthode pour détecter ces articles controversés. Le modèle commence par regrouper les articles similaires à un article concerné, en décrivant tous les articles par une représentation vectorielle moyenne de chacun des mots utilisés [Mik+13], puis en sélectionnant les plus proches grâce à une mesure de similarité cosinus. Un score basé sur l'analyse des réactions des utilisateurs sur chacun des articles Facebook est calculé, se basant sur les commentaires positifs et négatifs. Un score représentant la controverse présent dans les articles est aussi calculé, basé sur un dictionnaire de mots controversés. Enfin, ces deux scores sont combinés à un troisième, référençant l'intensité d'un article à partir du nombre de réactions, de commentaires et de partages liés à cet article et à ceux similaires.

Kim et Allan [KA19] font l'hypothèse que les sujets susceptibles de susciter des débats, avec de nombreux désaccords dans les médias en ligne, peuvent être considérés comme controversés. Les auteurs proposent une méthode non supervisée basée sur le contenu des articles. Leurs travaux se basent exclusivement sur un seul média en ligne ("the Guardian"), et les réactions proviennent directement du site lui-même. Un document est représenté par la combinaison de l'article et de ses commentaires. Deux modèles de langages sont entrainés, respectivement pour les articles et les commentaires, à partir de pseudo-labels calculés selon le nombre de désaccords dans les commentaires (score lui-même calculé à partir d'un réseau de neurones de type CNN). Le score de controverse représente la probabilité que le document ait plus de chances d'apparaître dans une collection de documents controversés ou non controversés, en se basant sur les probabilités *a posteriori* retournées par les modèles de langages.

Beelen, Kanoulas et Velde [BKV17] proposent une approche hybride, se basant sur les commentaires associés à un article, combinant les connaissances des sciences sociales et de l'informatique, afin de détecter un contenu controversé. Quatre types

de caractéristiques sont extraits pour chacun des articles : linguistique (variation du langage dans le débat), structurelle (nombre de commentaires, réponses, etc.), lexicale (calculé à partir d'un dictionnaire de mots pré-établi autour du désaccord), et émotionnelle (agrégation pour chaque commentaire d'un score retourné par un modèle pré-entrainé). Un modèle de type forêt aléatoire est ensuite appliqué à partir de ces caractéristiques, afin de prédire la controverse.

Les travaux présentés précédemment s'occupent de prédire la controverse avec une granularité correspondant à l'article. Rethmeier, Hübner et Hennig [RHH18] proposent une méthode multitâches supervisée, basée sur les CNNs, afin d'encoder les informations textuelles et prédire, au niveau des commentaires, la controverse des articles. Les commentaires, provenant d'un média en ligne [4], sont labellisés controversés ou non à partir des réactions des utilisateurs. Pour représenter les mots d'un commentaire sous forme de vecteurs numériques, et ainsi capturer les relations sémantiques et syntaxiques entre les mots, un algorithme de type "Word2Vec" est utilisé (pré-entrainé à partir de millions d'articles). Un CNN est ensuite entrainé sur deux tâches : une première basée sur la classification des commentaires à partir des mots représentés, et une seconde sur le genre de l'article.

Les méthodes présentées ici utilisent essentiellement des données textuelles et les commentaires associés aux articles afin de classifier les articles controversés sur les médias en ligne.

2.2.2 Analyse des médias sociaux

L'analyse de la controverse autour des médias en ligne reste une tâche compliquée, seul le point de vue de l'auteur/média étant mis en avant. L'utilisation progressive des médias sociaux (Twitter, Facebook, Reddit, etc.) apporte donc des points de vue et des opinions de plusieurs utilisateurs, avec donc une large quantité de données exploitables (notamment dans le cas de la controverse). Cela rend aussi son analyse plus complexe. Contrairement aux études présentées dans la section 2.2.1, les travaux présentés ici concernent l'analyse de la controverse autour des sujets dans leur globalité, et non autour des articles.

2.2.2.1 Détection et quantification de la controverse

a) Utilisation des caractéristiques structurelles (graphes). Plusieurs travaux ont été menés autour de la détection et la quantification de la controverse dans les médias sociaux, s'appuyant sur des caractéristiques et des définitions différentes de la controverse. Les approches d'analyse de la controverse basées sur l'information graphique reposent principalement sur trois étapes :

- La construction d'un graphe pour représenter une discussion sur un sujet en termes de nœuds utilisateur et d'interactions entre ces nœuds.
- Le partitionnement du graphe construit autour d'ensembles disjoints représentant probablement les communautés interagissant autour de la discussion.

4. *DerStandard.at*, journal autrichien.

- Le calcul de la mesure de controverse pour quantifier dans quelle mesure un sujet est controversé.

Ces approches étudient la controverse du point de vue de la polarisation des utilisateurs autour de différentes communautés. La figure 2.2 montre la séparation des utilisateurs autours de communautés pour les sujets dits controversés.

L'intuition derrière la polarisation des graphes de communautés a été proposée en premier par Conover et al. [Con+11], qui examinent la polarité de graphes utilisateurs de certains sujets politiques sur Twitter. Deux types de graphes sont analysés : le **graphe de retweets**, dans lequel les utilisateurs sont connectés si l'un d'entre eux a rediffusé un contenu produit par un autre, et le **graphe de mentions**, où les utilisateurs sont connectés si l'un d'entre eux a mentionné l'autre dans un message, y compris dans le cas de réponses à des tweets. Leur étude, basée sur l'analyse de la structure de ces graphes, à travers leurs communautés et leur homogénéité, démontre que le graphe de retweets présente une structure hautement modulaire, séparant les utilisateurs en deux communautés homogènes correspondant ici aux différents idéaux politiques.

En se basant sur cette intuition, Morales et al. [Mor+15] proposent une méthode pour quantifier la polarisation politique en utilisant Twitter comme source de données. En regardant la participation et le nombre de retweets des utilisateurs, les nœuds les plus influents sont labellisés grâce à la position politique exprimée dans leurs tweets. Des techniques de traitement du langage naturel, de l'anglais Natural Language Processing (NLP), sont utilisées pour classifier les positions politiques exprimées dans les tweets des influenceurs. On considère influenceurs, les utilisateurs influant l'opinion de leur audience sur les réseaux sociaux. Ces labels sont ensuite propagés à travers le graphe, afin de labelliser le reste des utilisateurs. Des mesures de polarisation sont proposées pour mesurer la polarisation du graphe, tels que l'indice de polarisation calculé à partir de la distribution d'opinions, ou encore de la centralité intermédiaire [Con+11]. Néanmoins, l'utilisation de labels générés à partir d'un petit échantillon considéré comme influenceurs (ne prenant seulement en compte le nombre de retweets), ainsi que l'application de la méthode sur un seul sujet controversé (et politisé), rend cette mesure peu fiable.

Pour répondre à ce problème, Garimella et al. [Gar+18a] proposent une autre méthode afin de quantifier la controverse sur Twitter, en utilisant une mesure basée sur la marche aléatoire, le RWC score ("Random Walk Controversy" en anglais), ainsi que le partitionnement du graphe en sous-graphes à l'aide de metis [KK95]. La méthode metis divise un graphe en plusieurs sous-graphes de taille équilibrée, en maximisant les connexions internes et en minimisant les connexions entre les sous-graphes. L'analyse de la controverse a été étudiée à partir de quatre types de graphes [Gar+16] : (1) un graphe de Retweet [5] construit afin de représenter quels utilisateurs sont d'accord sur un sujet donné, (2) un graphe de suivi qui se concentre sur le fait qu'un utilisateur suive le contenu d'un autre utilisateur sur les médias sociaux, (3) un graphe de contenu liant les utilisateurs partageant certains mots-clés et enfin (4) un graphe de sentiments reliant les utilisateurs partageant les mêmes opinions sur le sujet. L'étude

5. Action sur Twitter donnant la possibilité à un utilisateur de partager le contenu d'un autre utilisateur à son réseau.

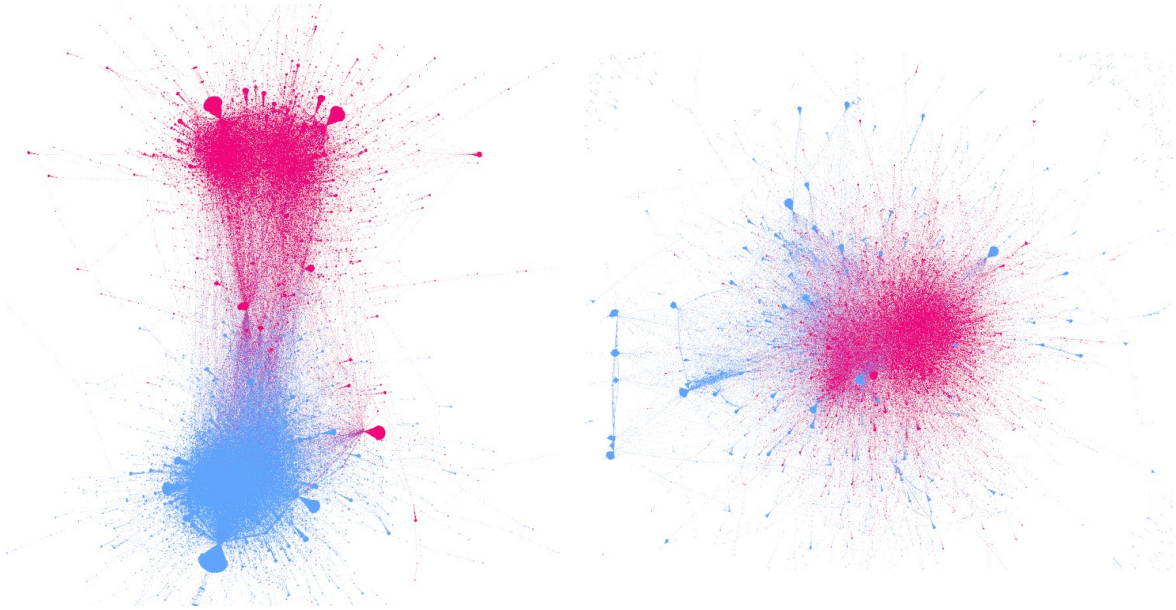

Figure 2.2 – *Gauche : Graphe de "Retweet" représentant la controverse autour du discours de Netanyahu en 2015. Droite : Graphe de "Retweet" représentant un sujet non controversé, la conférence "SXSW", autour de la musique, du cinéma et des médias interactifs, en juin 2015. Les deux graphes sont représentés en utilisant l'algorithme de projection ForceAtlas2 [Jac+14] pour la visualisation spatiale. Les couleurs représentent les labels des communautés calculées à partir de l'algorithme de partitionnement metis [KK95].*

montre l'utilité du graphe de retweets (Figure 2.2) par rapport aux autres graphes d'information. Cette approche est indépendante de la langue et du domaine et peut donc être appliquée facilement à n'importe quel sujet de discussion. Elle présente néanmoins l'inconvénient de ne pas exploiter d'autres attributs du graphe (comme le nombre de tweets par utilisateur, le nombre de followers, etc.). Cet inconvénient est pris en compte par exemple par EMAMGHOLIZADEH et al. [Ema+20] qui analysent en prenant compte des informations supplémentaires sur les nœuds et/ou les arêtes. La mesure de controverse [Gar+18a] est légèrement modifiée, en se basant sur des marches aléatoires biaisées pour prendre en compte toutes les informations ajoutées.

Il est connu que l'analyse de la controverse est confrontée au phénomène de "chambre d'écho". Il s'agit de situations dans lesquelles des personnes partageant les mêmes opinions renforcent mutuellement leur point de vue, car elles n'ont accès qu'à du contenu similaire à leurs opinions, et n'ont pas l'occasion d'être exposées à des points de vue opposés. Pour limiter ce phénomène, GARIMELLA et al. [Gar+18b] proposent une méthode pour réduire le score de controverse de la marche aléatoire (RWC) des sujets controversés. Pour cela, ils modifient la structure du graphe de Retweet, afin de réduire l'effet de polarisation du graphe. De nouvelles arêtes sont automatiquement ajoutées pour matérialiser les connexions entre les utilisateurs ayant des points de vue opposés. La réduction de la quantification de la controverse est définie ici comme un problème de recommandation d'arêtes. L'objectif est de calculer un nombre fixe de nouvelles arêtes entre les nœuds existants qui minimisent le score de controverse de la marche aléatoire. Les expérimentations menées montrent que l'ajout d'arêtes entre les nœuds ayant le degré de nœud le plus élevé entraîne la plus forte diminution du score de controverse. Cette approche permet de réduire l'impact des phénomènes de chambre d'écho d'une manière indépendante de la langue et du domaine, simplement en calculant les connexions entre utilisateurs de communautés différentes. Néanmoins, la méthode ne prend pas en compte des informations telles que le contenu des tweets, le profil de l'utilisateur, etc. La réduction de la controverse du point de vue de la polarité des communautés reste un défi ouvert dans les approches d'analyse de la controverse qui combinent des informations structurelles et textuelles.

La plupart des approches d'analyse des controverses basées sur les graphes supposent que les discussions polarisées sont caractérisées par la qualité du partitionnement des communautés (modularité). GUERRA et al. [Gue+13] rejettent cette hypothèse et considèrent que la métrique traditionnelle de polarisation autour de la modularité n'est pas nécessairement une métrique suffisante, puisque les réseaux non (ou peu) polarisés peuvent également être divisés en communautés. Une nouvelle mesure de polarisation, basée sur la frontière entre les communautés, est définie pour mieux différencier les graphes polarisés et non polarisés. Les auteurs démontrent empiriquement que les communautés polarisées se retrouvent souvent avec peu de nœuds de haut degré le long de la frontière.

b) Utilisation des caractéristiques textuelles. D'autres travaux se basent sur l'analyse seule des textes [Ras+21]. Ces recherches permettent notamment de mieux appréhender la controverse, par le biais de l'analyse de sentiments, de la polarité ou de la prise de position des utilisateurs concernés.

En se focalisant sur le réseau social Twitter, Addawood et al. [Add+17] se concentrent sur un degré de granularité moindre, détectant les tweets associés à des sujets controversés ou non controversés et non les sujets en eux-mêmes, afin de mieux comprendre les caractéristiques de langage lié à la controverse. Afin de représenter chaque tweet, et à partir de méthodes provenant de l'état de l'art, différents types de caractéristiques sont extraits à partir du contenu :

1. **Caractéristiques emphatiques** : quantification du recours au langage emphatique, représenté par l'orthographe, la ponctuation ou encore le lexique (intuition dans laquelle les tweets appartenant à des sujets controversés ont recours davantage à ce style de langage) ;
2. **Caractéristiques de langage** : caractéristiques représentant un point de vue psychologique [Boy+22] et grammatical ;
3. **Caractéristiques spécifiques à Twitter** : attributs des tweets, tels que les URLs, les mentions ou les hashtags utilisés, représenté par leur nombre d'occurrences dans le tweet.

En se basant sur ces caractéristiques textuelles, différents modèles d'apprentissage automatique (SVM, arbre de décision, classifieur bayésien naïf) sont entrainés afin de classifier l'appartenance des tweets à des sujets controversés ou non.

Cependant, la granularité ici ne permet pas de prédire si le sujet en lui-même est controversé ou non. Rashed et al. [Ras+21] proposent dans leur approche d'analyser la controverse autour de la détection de prise de position dans les sujets controversés, la controverse étant aussi étudiée du point de vue de la polarité des utilisateurs. La détection de prise de position vise à identifier la position (favorable, défavorable, neutre, etc.) d'un utilisateur à l'égard d'une entité donnée (sujet, personne, etc.). Un sous-ensemble d'utilisateurs est étiqueté selon leurs positions. Ces étiquettes sont obtenues manuellement pour certains utilisateurs, les plus actifs, et automatiquement pour d'autres en utilisant une méthode de propagation de labels autour des retweets émis par les utilisateurs. Les utilisateurs étiquetés et leurs différents types de caractéristiques sont utilisés pour former des modèles de classification (SVM et FastText) afin de prédire les positions des utilisateurs non étiquetés. Les positions des utilisateurs sont utilisées pour regrouper les utilisateurs avant d'appliquer des mesures de controverse et ainsi quantifier la controverse. Exploitant des techniques récentes de classification et de propagation de labels, cette approche présente néanmoins l'inconvénient de devoir étiqueter manuellement certains utilisateurs et de favoriser les utilisateurs actifs par rapport aux utilisateurs non actifs.

En se basant sur la même polarisation des communautés, Zarate et al. [Zar+20] proposent de quantifier la controverse dans les médias sociaux par l'analyse du contenu en utilisant de larges modèles de langage récents tel que BERT [Dev+19]. Ce modèle est basé sur le concept des Transformers et des mécanismes d'auto-attention. Après avoir créé un graphe de retweets et partitionné ce graphe en plusieurs communautés, les tweets des deux plus grosses communautés d'utilisateurs sont labellisés avec le même label que leur communauté. Puis, un modèle BERT est entrainé à partir de ces tweets, afin de récupérer un vecteur de représentation de chacun des tweets. À partir de ces

vecteurs, un score de controverse est calculé pour chaque cluster. Ce score correspond à la distance de leur centroïde avec les utilisateurs centraux, comparé ensuite à la distance globale. Intuitivement, plus la distance entre les deux clusters est grande, plus les tweets des communautés seront considérés comme différents, et donc le sujet plus polarisé et controversé. Même si certaines approches [Ras+21 ; Zar+20] utilisent le graphe afin de montrer l'importance que peut avoir le contenu dans l'analyse de la controverse, les caractéristiques structurelles en elles-mêmes ne sont pas prises en compte, provoquant une perte non négligeable d'informations.

Pour palier les limites des deux familles de méthodes précédentes, plusieurs approches hybrides ont été proposées, combinant à la fois le contenu et la structure des sujets.

Zarate et Feuerstein [ZF20] étendent leur approche [Zar+20] basée sur le contenu afin de quantifier la controverse, et utilisent une mesure de controverse basée sur la propagation des probabilités d'appartenance d'un utilisateur à une communauté dans le graphe de retweets. Ils considèrent tous les utilisateurs (y compris ceux n'ayant pas publié de tweets). Un graphe de retweets est créé, puis partitionné à l'aide de méthodes basées sur la structure du graphe en plusieurs clusters. Seules les deux plus grandes communautés d'utilisateurs représentées sont conservées. Chaque utilisateur est représenté par la concaténation de ses tweets, et est labellisé par sa communauté. Un modèle TAL est entrainé afin de classifier les utilisateurs et de récupérer leurs représentations vectorielles. Seuls les utilisateurs ayant une probabilité très forte selon le modèle d'appartenir à une des communautés sont labellisés en tant que tels, le reste des utilisateurs est labellisé selon une méthode de propagation de labels. Une mesure de controverse est calculée à partir du moment dipolaire, appliqué en physique sur les molécules, afin de mesurer la séparation des charges électriques positives et négatives, indiquant ainsi la polarité de celle-ci.

Al-Ayyoub et al. [Al-+18] proposent de compléter les travaux de Garimella et al. [Gar+18a] en ajoutant à l'analyse structurelle du graphe (à partir du RWC score) une mesure d'analyse de sentiments à partir du texte, mesurant la polarité des sentiments des tweets (positive, neutre, négative). Une grande polarité des tweets indique alors un haut score de controverse.

Mendoza, Parra et Soto [MPS20] soulignent le fait que la dynamique d'un réseau social est caractérisée par deux facteurs principaux : les utilisateurs qui interagissent entre eux et les entités impliquées dans une discussion donnée. Ils exploitent ensuite les commentaires des utilisateurs sur les entités nommées pertinentes qui apparaissent dans le texte pour déduire la nature de la tendance (positive, négative, neutre) de ces utilisateurs à l'égard des entités nommées. Ils proposent une méthode pour générer des réseaux d'utilisateurs conditionnés par la relation entre l'utilisateur et l'entité nommée. Le graphe d'utilisateurs obtenu permet d'analyser les interactions entre les utilisateurs ayant des points de vue opposés sur les entités nommées. La controverse est prédite par différentes mesures, dont le RWC présenté par Garimella et al. [Gar+18a]. Les auteurs montrent que la détection des controverses est améliorée lorsque des entités nommées polarisées sont utilisées.

D'autres approches se focalisent sur les interactions et les débats entre utilisateurs

des différents camps, plutôt que sur la polarisation des utilisateurs. Certains travaux analysent les arbres de commentaires du réseau social Reddit [6] sous différents posts, appartenant à des catégories (ou "subreddits"). Un commentaire appartient à un post Reddit, qui lui-même appartient à un subreddit. Le graphe Reddit (l'arbre de commentaires) est ensuite construit. Chaque nœud correspond à un texte (post ou commentaire). Une arête (ou lien) entre deux nœuds existe lorsqu'un texte commente un autre.

Hessel et Lee [HL19] démontrent que le fait de combiner des caractéristiques structurelles (nombre de commentaires, rapport profondeur maximale/total des commentaires, profondeur moyenne des nœuds, etc.) d'un arbre de commentaires d'une discussion Reddit avec des caractéristiques textuelles produites par des modèles de langage tels que BERT [Dev+19] peut améliorer les performances prédictives de la détection précoce de la controverse au niveau d'un post Reddit. Les auteurs caractérisent chacun des posts à partir de trois ensembles de variables :

1. **C-RATE** : score représentant les métadonnées de la discussion (nombre de commentaires, temps d'attente avant le premier commentaire, etc.) ;
2. **C-TREE** : score représentant les aspects structurels de l'arbre de discussion (profondeur de l'arbre, proportion de commentaires au premier niveau, moyenne de profondeur des nœuds, etc.) ;
3. **C-TEXT** : score encodant chaque commentaire dans un vecteur de représentation, extrait depuis un modèle de langage BERT, puis les agrégeant pour chaque post.

Plusieurs modèles d'apprentissage automatique sont ensuite comparés afin de prédire les posts controversés. L'approche est expérimentée sur six jeux de données différents (provenant de six "subreddits") et sur différentes périodes de temps, contenant chacun un nombre de posts labellisés controversés ou non. Ce label est directement calculé à partir de différentes mesures provenant de caractéristiques de chacun des posts complets (nombre de votes positifs et négatifs).

Le texte est donc représenté ici par le biais de réseau de neurones profonds, le rendant plus complexe et précis, ce qui n'est pas le cas des caractéristiques du graphe, représentées par des valeurs statistiques. Zhong et al. [Zho+20] adoptent une approche ayant pour objectif d'encoder le graphe à l'aide des réseaux de neurones graphiques ("Graph Neural Networks" en anglais). Les auteurs représentent la détection de posts controversés comme une tâche de classification de graphes, et s'appuient sur des réseaux de neurones graphiques à convolution (GCN). Les auteurs visent à intégrer les informations extraites de la structure de l'arbre des commentaires ainsi que le contenu de l'article et de ses commentaires. Ils affirment également que l'exploitation des seules caractéristiques sémantiques et structurelles des commentaires attachés au post n'est pas suffisante, et proposent d'exploiter les articles connexes sur le même sujet ("subreddit"). L'opération de convolution est utilisée pour générer une représentation vectorielle pour chaque nœud. Ces derniers sont agrégés ensuite pour représenter le graphe sous forme de vecteur. Enfin, une couche de classification détermine alors si cette représentation du graphe est controversée ou non. Le modèle est entrainé en

6. plateforme sociale de discussion en ligne

parallèle sur plusieurs tâches afin de séparer les caractéristiques liées et non liées aux différents posts, pour la détection des posts en prenant compte des subreddits. Cette méthode présente certaines limites. La structure de l'arborescence des commentaires d'un article ignore complètement l'utilisateur écrivant un texte (post ou commentaire), et n'inclut pas d'autres types de comportement de l'utilisateur comme l'approbation d'un commentaire, comme le fait le Retweet sur Twitter. De plus, la stratégie de représentation des graphes est simple (agrégation de tous les commentaires) et considère tous les textes (nœuds) au même niveau. Enfin, l'utilisation de relations entre les posts pourrait interférer avec la tâche principale de détection. Néanmoins, cette approche est la première étude qui se concentre sur les GNN afin d'analyser la controverse, ce qui a été longuement étudié lors de nos travaux. Toutes ces approches montrent que la structure du graphe n'est pas suffisante pour quantifier le niveau de controverse et que l'utilisation d'informations supplémentaires tels que le contenu peut s'avérer utile.

2.2.2.2 Explication et visualisation des sujets controversés

Définir et comprendre les sujets controversés reste une tâche assez complexe. L'explicabilité de la controverse consiste à définir les sujets controversés, soit en expliquant et en analysant les diverses opinions autour de ces sujets, soit en essayant de caractériser ce que représente la controverse [HL19].

Plusieurs approches se concentrent sur le contenu diffusé sur les réseaux sociaux afin de mieux comprendre ces sujets controversés. Kim et Allan [KA19] proposent une méthode expliquant quel sujet est controversé dans un document à l'aide de leur classifieur de controverse basé sur le contenu. La controverse est expliquée en générant des phrases qui décrivent le sujet controversé dans le document. Les auteurs se concentrent ici sur des articles de médias en ligne contenant à la fois le contenu de l'article et les commentaires des utilisateurs sur l'article.

Koncar, Walk et Helic [KWH21] proposent une analyse statistique des caractéristiques textuelles pour les besoins d'explication de la controverse. Ils visent à mesurer les forces prédictives des caractéristiques du texte individuellement pour la controverse sur la plateforme Reddit, comme indiqué dans la figure 2.3. La méthode se focalise spécifiquement sur certains aspects, tels que l'utilisation de mots-clés, le style d'écriture, l'analyse de sentiments ou l'implication des utilisateurs. Les résultats de l'approche montrent clairement que la plupart des caractéristiques reflètent la controverse de manière similaire dans toutes les langues.

D'autres approches se focalisent sur l'explication des communautés polarisées autour des sujets controversés, en synthétisant le contenu important dans chacune d'entre-elles. Guo et al. [Guo+15] présentent dans leurs travaux une méthode afin de résumer les opinions contradictoires des utilisateurs dits "experts" dans le corpus de tweets reliés au sujet en question. Le modèle intègre les avis d'experts et les avis ordinaires et produit une liste de paires d'arguments représentatifs et contrastés pour le sujet controversé. Un algorithme semi-supervisé de regroupement d'arguments est appliqué ensuite, afin de regrouper les arguments "positifs" ou "négatifs" vis-à-vis d'un sujet. Cependant, il est difficile de seulement considérer l'aspect sentiment (positif ou négatif) pour représenter la controverse.

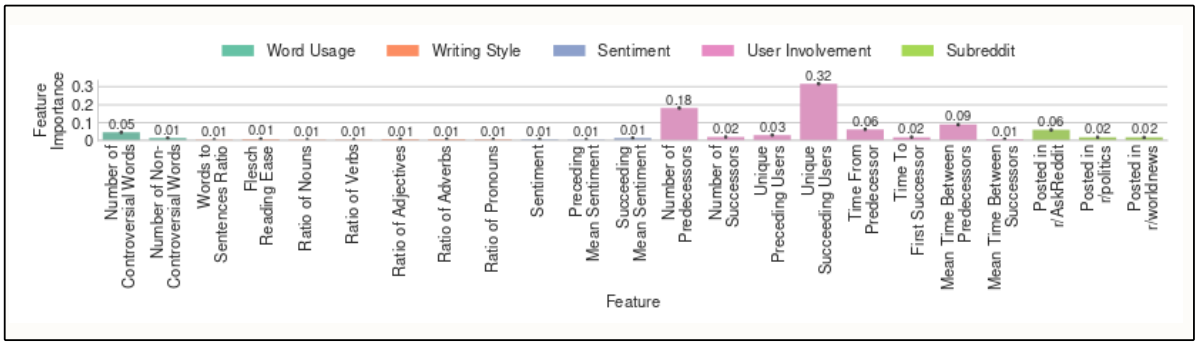

Figure 2.3 – *Visualisation des caractéristiques du texte les plus importantes lors de la prédiction de la controverse autour des sujets. Chaque caractéristique est triée par catégorie (indiquée par des couleurs différentes). L'image provient des travaux de Koncar, Walk et Helic [KWH21].*

Jang et Allan [JA18] s'intéressent aux sujets controversés en résumant les différentes prises de position des utilisateurs de chacune des communautés extraites du graphe de retweets. Un classifieur est utilisé de manière non supervisée, afin de repérer les tweets qui résument le mieux les positions de chaque communauté de sujets controversés. Les auteurs caractérisent les tweets sélectionnés pour chaque communauté par trois scores, calculés au préalable, utilisés afin de classer ces tweets :

1. **Degré de prise de position** : Si la prise de position est forte ou non ;
2. **Articulation** : Si le tweet est clair, persuasif, logique et rédigé dans un langage approprié ;
3. **Pertinence du sujet** : Si le tweet est explicite et pertinent dans le contexte du sujet.

Malheureusement, ces approches limitent l'explicabilité de la controverse à la présentation d'arguments ou de prises de positions de part et d'autre d'un débat contradictoire, représentant une définition partielle de la controverse.

Enfin, d'autres approches se concentrent sur la compréhension de la structure des relations entre utilisateurs. Dans leur approche pour quantifier la controverse sur Twitter, Morales et al. [Mor+15] projettent une explication visuelle d'un sujet controversé d'un point de vue temporel et géographique. Ces projections visuelles, mises en avant dans la figure 2.4, permettent une possible interprétation des points de vue selon des communautés géographiques, comme celles d'événements particuliers au cours du temps.

Les techniques d'analyse graphique sont également exploitées par Coletto et al. [Col+17] pour analyser les interactions entre les utilisateurs et rechercher des motifs locaux qui caractérisent les sujets controversés. Les auteurs visualisent différents graphes d'interactions entre utilisateurs sur Twitter focalisés sur les réponses (commentaires) autour des posts. L'aspect temporel est aussi visualisé, comme montré dans la figure 2.5, afin de voir l'évolution des mesures de controverse autour d'un sujet à travers le temps.

Toutes ces approches ont pour but d'aider à la compréhension de la controverse, soit par l'analyse du contenu et des communautés, soit en analysant la structure des

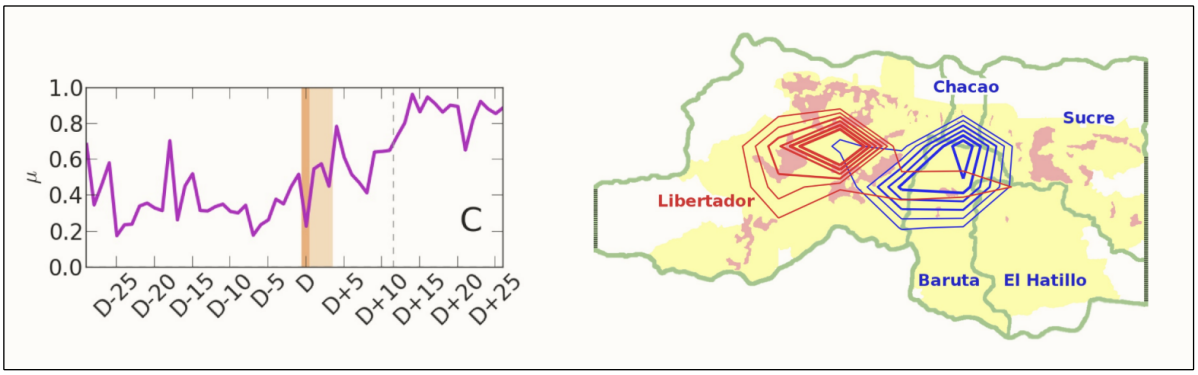

FIGURE 2.4 – *Visualisation de l'information controversée d'un sujet d'un point de vue temporel et géographique. À gauche, l'évolution du score de controverse d'un sujet est annoté au cours du temps. À droite, la polarisation des communautés de la ville de Caracas (Venezuela) est représentée, en utilisant des fonctions de densité sur la probabilité d'un tweet d'appartenir à un utilisateur d'une communauté. Les images proviennent des travaux de* MORALES *et al.* [Mor+15].

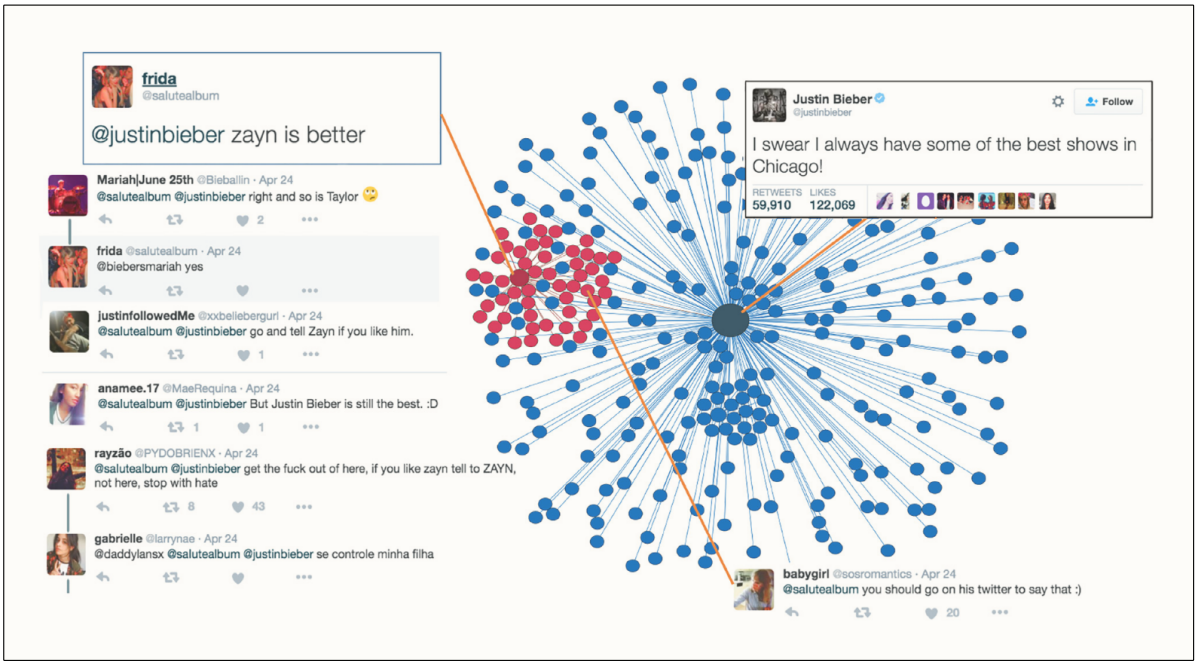

FIGURE 2.5 – *Visualisation du contenu des utilisateurs de chaque communauté, à partir du graphe construit. L'image provient des travaux de* COLETTO *et al.* [Col+17].

relations entre utilisateurs. L'étude de la structure des interactions entre utilisateurs peut donc être poussée à l'aide des réseaux de neurones graphiques, de l'anglais Graph Neural Networks (GNN), présentés dans la section 2.3. La figure 2.6 permet de visualiser une représentation globale des principales études menées autour de l'analyse de la controverse, selon les différentes méthodes utilisées.

Dans le cadre des travaux de cette thèse, nous nous positionnons sur l'analyse de la controverse sur les médias sociaux. Nous étudions deux types de médias en particulier, Twitter et Reddit. Comme évoqué dans la section 2.2.2, plusieurs travaux se sont focalisés sur les informations structurelles (graphes utilisateurs, arbres de commentaires) et textuelles (tweets, posts, commentaires) afin d'analyser la controverse. Nos approches autour de la détection et quantification de la controverse se baseront notamment sur l'utilisation hybride de ces caractéristiques, et de l'application des Réseaux de neurones graphiques (GNN) présentés dans la section 2.3.

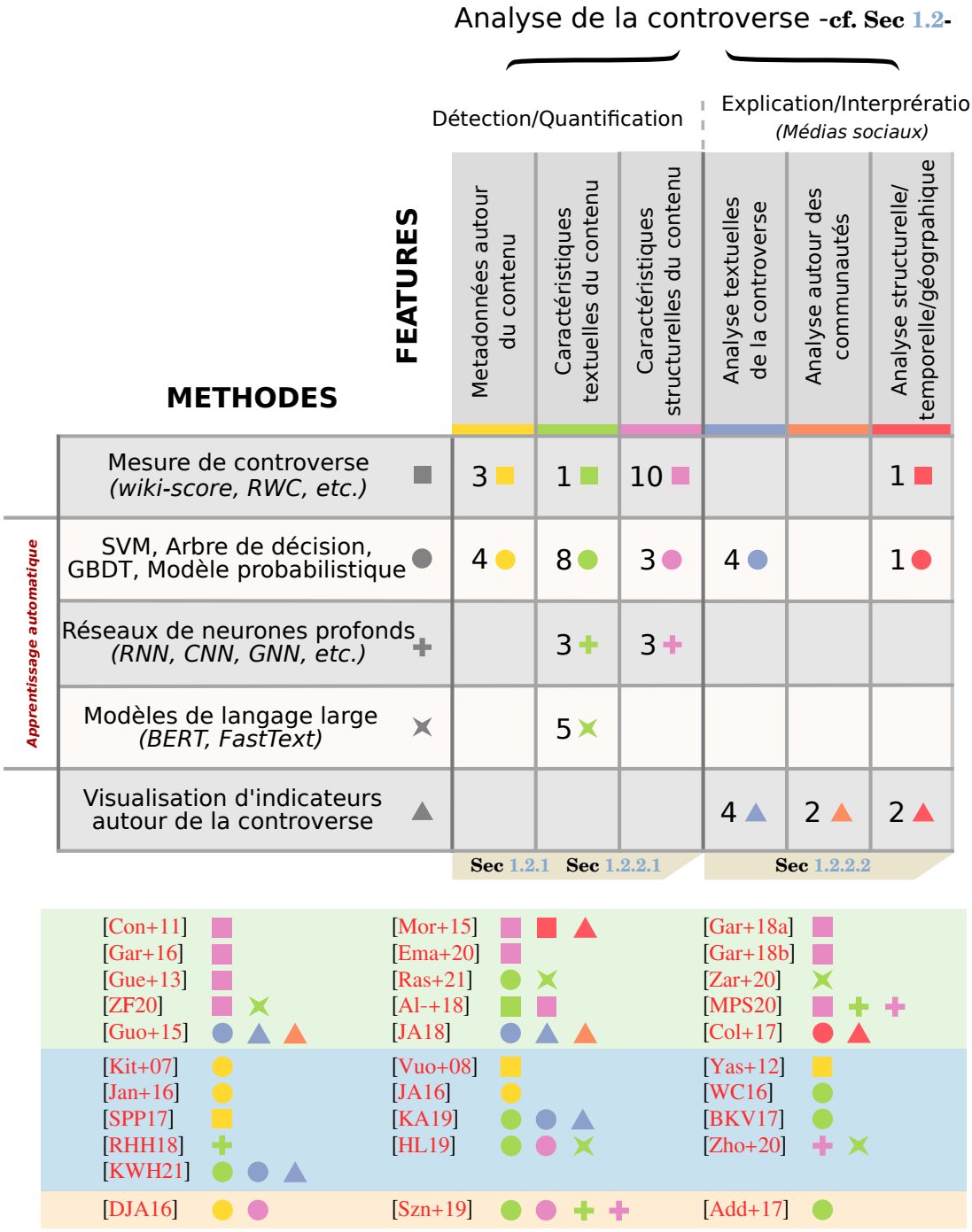

FIGURE 2.6 – *Tableau dynamique représentant différents travaux autour de la controverse, selon les caractéristiques et les méthodes utilisées. Certaines approches peuvent appartenir à plusieurs catégories, qui sont explicitées pour chaque approche dans la légende. La légende est séparée en trois catégories, représentant la caractéristique principale utilisée pour définir la controverse : En vert, les méthodes utilisent l'aspect polarisé de la controverse et le rassemblement des utilisateurs autour de communautés. En bleu, les méthodes représentent la controverse autour des litiges entre utilisateurs autour des sujets controversés, pouvant être présents lors de discussions/interactions sur les médias sociaux, ou de guerre de modifications des articles ("edit wars" en anglais) sur Wikipédia par exemple. En orange, les méthodes se consacrent purement au contenu, à l'analyse de sentiments, ou d'entités relatives à la controverse.*

2.3 Représentation des données sous forme de graphes et application sur des réseaux de neurones (GNN)

La représentation des données sous forme de graphes est très utilisée dans le monde moderne. Elle permet de modéliser des informations sous forme de relations entre objets. Les graphes peuvent représenter des situations variées, simples, comme les molécules chimiques [Yin+18], des structures de données (arbres, listes chaînées) ou plus complexes, tels que les réseaux de communications (flux de données), les réseaux de connaissances (système d'information), les réseaux de transport [JL22] ou encore les réseaux sociaux [Gar+18a ; HL19]. Les objets sont représentés par des nœuds (ou sommets) et les relations par des liens (ou arêtes). On distingue ici les graphes selon différentes caractéristiques :

1. **Orientation du graphe.** Un graphe peut être orienté, avec une direction pour un lien entre deux nœuds, ou non orienté ;
2. **Homogénéité du graphe.** Dans un graphe homogène, tous les nœuds et liens sont du même type, contrairement aux graphes hétérogènes, qui peuvent avoir des nœuds et/ou arêtes avec différents types ou attributs ;
3. **Temporalité du graphe.** Un graphe est statique si les données d'entrée ne varient pas au cours du temps. Dans le cas contraire, le graphe est considéré comme dynamique.

Le principal intérêt du graphe est qu'il permet de représenter des données non structurées. Contrairement aux images, un graphe peut être de taille variable, présentant des nœuds non ordonnés et avec un nombre variable de voisins. Une image peut être considérée comme un graphe de taille fixe, avec des nœuds ordonnés (du 1er au dernier pixel). Pour les réseaux de neurones classiques, cette non-structuration des données en entrée représente un vrai défi, alors que ces derniers entrainent des matrices de poids de tailles fixes et ordonnées.

L'apparition progressive des réseaux de neurones graphiques au cours des années 2010 [PAS14 ; HBL15 ; KW17], aussi appelés Graph Neural Networks ou GNN, tend à résoudre ce problème de représentation des données non structurées. Les réseaux de neurones graphiques sont une famille de modèles d'apprentissage automatique qui traitent des données organisées sous forme de graphes en utilisant des techniques d'apprentissage profond (ou Deep Learning) pour effectuer des tâches reliées au graphe de bout en bout. Elles permettent notamment de représenter les caractéristiques des nœuds (ou d'un graphe) sous la forme d'un vecteur de plongement dans un espace euclidien (ou embedding), pouvant ensuite être utilisé dans des réseaux de neurones classiques. Ces tâches regroupent de manière non exhaustive la classification de nœuds ou de graphes [KW17 ; HYL17 ; Vel+18], la prédiction de liens [Xie+22] ou encore la génération de graphes [Wu+21]. Elles permettent notamment l'identification de certaines protéines [ZX20 ; Yin+18], des fausses informations circulant sur les réseaux sociaux [Ham+20], la prédiction du trafic dans une ville ou encore l'optimisation des systèmes de recommandation [Wu+21]. Les réseaux de neurones graphiques peuvent être divisés en plusieurs catégories : les modèles de noyau, les modèles récurrents ainsi

que **les modèles à convolution**. Dans cet état de l'art, nous portons une attention particulière à cette dernière catégorie, à laquelle nous nous intéressons dans les chapitres 4 et 5.

Les méthodes de noyau (ou "kernel" en anglais) [Wu+21] sont basées sur la théorie des noyaux, qui permet de définir des fonctions de similarité non linéaires entre les nœuds d'un même graphe. Ces fonctions sont donc déterministes, et définies en termes de produits scalaires dans un espace de représentation de grande dimension. En utilisant cette formulation, les GNN peuvent capturer des relations de voisinage complexes dans les graphes. Concernant les réseaux de neurones récurrents appliqués aux graphes [Wu+21], ces modèles consistent à propager les informations dans le graphe à travers des cellules récurrentes. Ces cellules prennent en compte les informations des nœuds voisins et mettent à jour les représentations des nœuds à chaque étape de la propagation. Ainsi, les GNN prennent en compte les dépendances à long terme entre les nœuds.

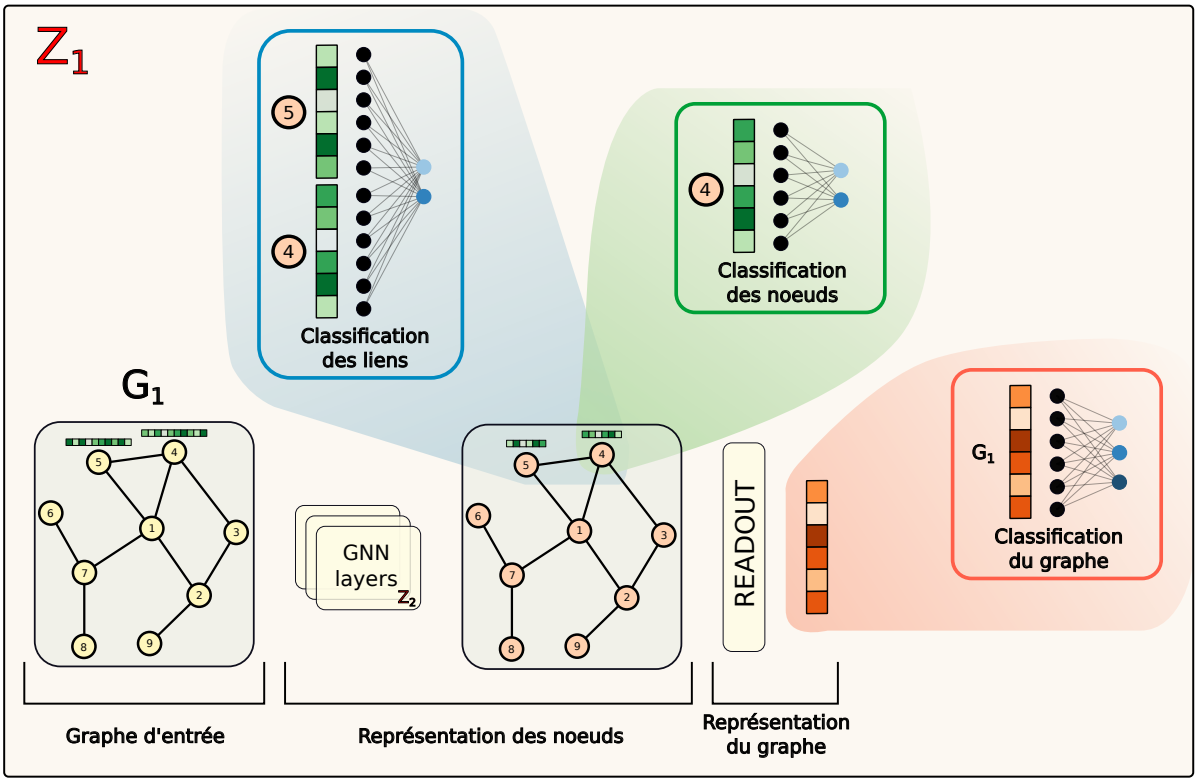

FIGURE 2.7 – *Schéma résumant le fonctionnement général des réseaux de neurones à convolutions. Le bloc Z_2 "GNN layers" représente les couches cachées du réseau (développé dans la figure 2.9), et la fonction "READOUT" la fonction d'agrégation (ou de pooling) du graphe. Les trois tâches d'apprentissage (semi-)supervisées sont représentées : la classification des liens dans l'**encadré bleu**, des nœuds dans l'**encadré vert** et du graphe dans l'**encadré rouge**.*

Les convolutions, appliquées à des réseaux de neurones profonds, permettent d'appliquer des filtres sur les données d'entrée afin de mettre en évidence certaines caractéristiques. Utilisées d'abord dans le traitement des images avec les CNN [Lec+98], elles ont ensuite été généralisées pour répondre à différents besoins, comme le traitement automatique du langage ou du signal [WHD+18], et donc du traitement des

graphes [KW17]. Dans le cas des graphes, le filtre permet de prendre en compte la représentation du voisinage lors du calcul de la nouvelle représentation des nœuds dans chaque couche l du réseau de neurones. Cela permet notamment, par le contrôle du nombre de couches cachées, de retenir une information plus locale ou plus globale du graphe dans la nouvelle représentation des nœuds. Les réseaux de neurones à convolutions peuvent être utilisés pour deux types de tâches différentes : **La représentation des nœuds**, qui peut ensuite être suivie par **la représentation du graphe**. La figure 2.7 schématise l'organisation de ces deux tâches, ainsi que les tâches de classification supervisées existantes.

2.3.1 Formalisation et notation

Afin de simplifier la lecture de ce chapitre, nous allons considérer des graphes, orientés ou non, homogènes, statiques. Les liens entre les nœuds ne contiennent pas de caractéristiques. Différentes approches ont abordé la question des graphes hétérogènes et des graphes dynamiques, mais elles ne seront pas examinées dans cette section. Un graphe est représenté par $G = (V, E)$, avec V l'ensemble des nœuds et E l'ensemble des liens entre nœuds. $v_i \in V$ désigne un nœud, et $e_{ij} = (v_i, v_j) \in E$ désigne un lien allant de v_i à v_j. L'ensemble $N(v) = \{u \in V, si(v, u) \in E\}$ représente l'ensemble des voisins du nœud v. A représente la matrice d'adjacence, de taille $n \times n$ avec n le nombre de nœuds. Les attributs des nœuds sont regroupés dans une matrice $X \in \mathbb{R}^{n \times d}$, avec d le nombre d'attributs.

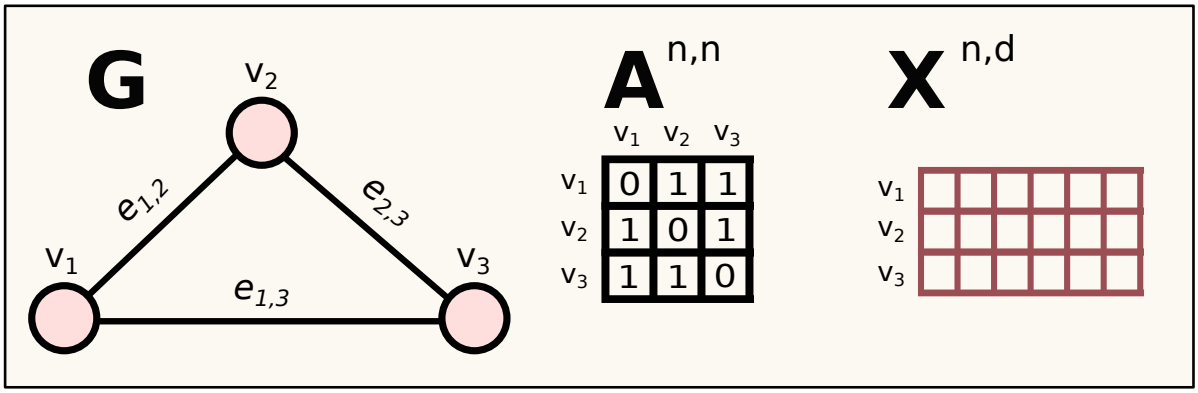

Figure 2.8 – *Exemple générique d'un graphe G, représenté par ses nœuds et liens, ainsi que sa matrice d'adjacence A et sa matrice de caractéristiques X.*

Les notations utilisées sont répertoriées dans la table 2.1. La figure 2.8 représente un exemple générique de graphe, ainsi que ces matrices correspondantes.

2.3.2 Représentation des nœuds

L'un des objectifs des réseaux de neurones (GNN) graphiques est de pouvoir représenter ces composantes dans un sous-espace de représentation, en prenant en compte non-seulement ces caractéristiques, mais aussi celles du graphe auquel il appartient. Les réseaux de neurones à convolution permettent de prendre en compte des localités par le biais des filtres, et donc du voisinage lors du calcul de la nouvelle

TABLE 2.1 – *Notations utilisées pour la représentation des graphes*

Notation	Description
G	Graphe
V	Ensemble de nœuds
E	Ensemble de liens (arêtes)
v	Nœud $v \in V$
e_{ij}	Lien (arête) entre les nœuds i et j, $e_{ij} \in E$
$\mathcal{N}(v)$	Ensemble de voisins du nœud v
$\tilde{\mathcal{N}}(v)$	Ensemble de voisins du nœud v, y compris v
A	Matrice d'adjacence
D	Matrice des degrés de la matrice A
n	Nombre de nœuds, $n = \|V\|$
m	Nombre de liens, $m = \|E\|$
d	Dimension du vecteur de caractéristiques des nœuds
b	Dimension de la couche cachée d'un réseau de neurone
$X \in \mathbb{R}^{n \times d}$	Matrice des caractéristiques du graphe d'entrée
$H^l \in \mathbb{R}^{n \times b}$	Matrice des caractéristiques des nœuds du graphe à la couche l
$h_v^l \in \mathbb{R}^b$	Représentation du nœud v à la couche l
W^l	Matrice de poids entrainable de la couche cachée l
l	Index de la couche cachée du réseau
σ	Fonction d'activation de la couche du réseau de neurone

représentation. Le nombre de couches utilisées permet de décider l'ordre du niveau de localité. Comme le montre la figure 2.9, plus le nombre de couches est élevé, plus la représentation des nœuds est globale, et inversement. Ces réseaux à convolutions peuvent se baser sur deux théories : les méthodes spectrales [Bru+14 ; DBV16 ; KW17 ; ZM18] et spatiales [Xu+19 ; HYL17 ; Vel+18 ; Zha+18a ; Gil+17 ; Sch+18].

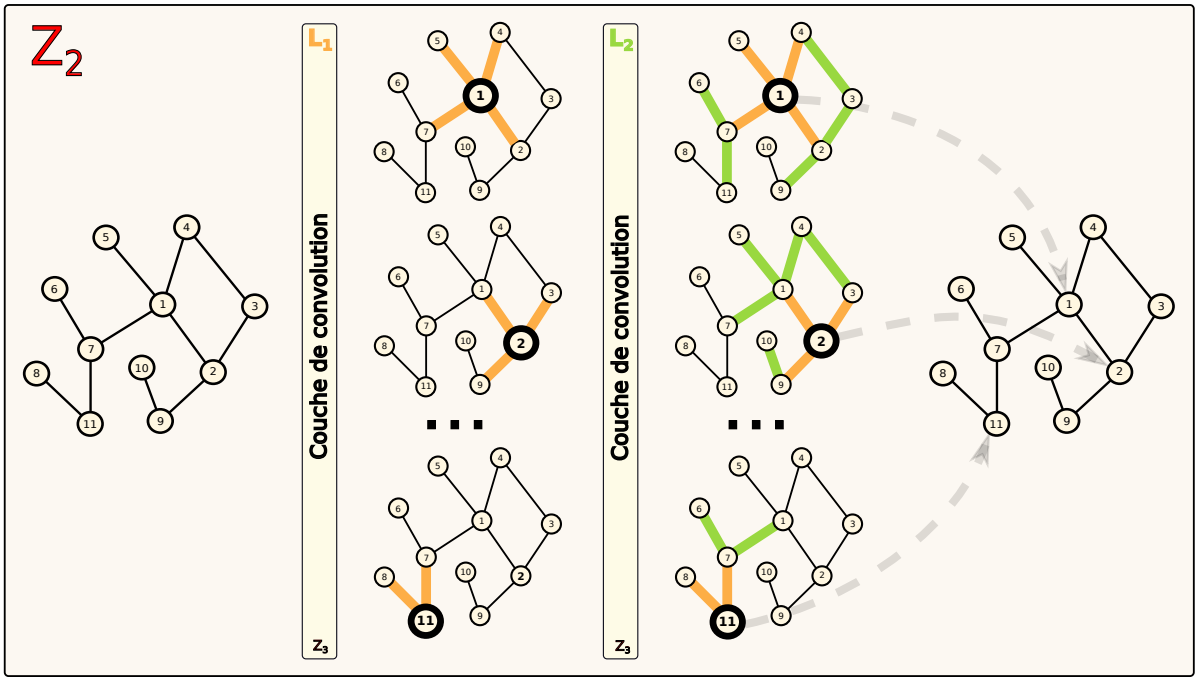

Figure 2.9 – *Schéma résumant le fonctionnement général d'un réseau de neurones graphiques à convolution à 2 couches. Le modèle retranscrit l'information structurelle du graphe, via l'apprentissage de nouvelles représentations des nœuds en utilisant les nœuds voisins de 1er niveau après la couche L1 et jusqu'au 2ème niveau après la couche L2*

.

2.3.2.1 Théorie spectrale

La théorie spectrale permet d'analyser les propriétés spectrales des signaux sur les graphes. L'opération de convolution correspond alors à l'élimination des bruits des signaux graphiques [Wu+21]. Elle se base sur les éléments spectraux (matrice d'adjacence, laplacienne, valeur propres) ainsi que la transformation de Fourrier [KW17] afin d'analyser les propriétés du graphe et ainsi définir un filtre sur le graphe. Différentes méthodes ont été proposées afin de définir ce filtre. Les GNN supposent notamment que le filtre peut être considéré comme un ensemble de paramètres pouvant être appris, prenant en compte les signaux graphiques sur différents canaux.

Bruna et al. [Bru+14] (Spectral CNN) utilisent la matrice des vecteurs propres classés par valeurs propres de la matrice Laplacienne, accompagnée d'une matrice diagonale de paramètre d'apprentissage. Cette solution suscite certains problèmes. La décomposition en vecteurs propres de la matrice Laplacienne, rend les filtres dépendants au graphe même, et donc l'impossibilité de les appliquer sur des graphes de différentes structures. De plus, la décomposition augmente la complexité de calcul en $O(n^3)$.

DEFFERRARD, BRESSON et VANDERGHEYNST [DBV16] (ChebNet) appliquent plusieurs approximations et simplifications de ce filtre pour réduire ce problème de complexité en $O(m)$, en utilisant notamment les polynômes de Tchebychev. Ces nouveaux filtres, définis aussi dans l'espace, permettent notamment d'extraire des caractéristiques locales indépendamment de la taille du graphe, indiquant un premier pas vers la théorie spatiale.

KIPF et WELLING [KW17] (GCN) proposent une nouvelle approximation de ChebNet [DBV16] pour éviter le sur-apprentissage et restreindre le nombre de paramètres d'apprentissage. Les auteurs utilisent la matrice d'adjacence du graphe dans l'équation de propagation vers l'avant du réseau de neurones, représentée par l'équation 2.1 afin de tenir compte des propriétés structurelles du graphe. Cette méthode est représentée dans la figure 2.10.

$$H^{l+1} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{l}W^{l}\right) \tag{2.1}$$

Dans l'équation 2.1, H^{l+1} représente la sortie du graphe après la couche l, H^l les caractéristiques des nœuds en entrée de la couche l, W^l la matrice de poids, σ la fonction d'activation, $\tilde{A} = A + I_n$, avec I la matrice d'identité et $\tilde{D}$ sa matrice respective des degrés. $\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ représente finalement la matrice d'adjacence normalisée par ces degrés, utilisée pour éviter la production d'instabilités numériques. À noter qu'à la première couche $l = 0$, $H^0 = X$, avec X la matrice des caractéristiques des nœuds du graphe d'entrée. Cette méthode est souvent considérée comme une avancée majeure pour les réseaux de neurones à convolutions graphiques, faisant aussi le pont vers la théorie spatiale.

D'autres travaux ont ensuite essayé d'améliorer cette approche. ZHUANG et MA [ZM18] proposent une approche (DGCN) pour encoder à la fois l'information locale et globale sans utiliser plusieurs couches, en parallélisant deux GCN [KW17].

Comme expliqué auparavant, le problème majeur de cette théorie réside dans le fait qu'elle se base sur les propriétés d'un graphe fixe. En effet, le modèle opère sur l'ensemble du graphe pour calculer les nouvelles représentations. Le graphe et tous les états intermédiaires sont stockés, ce qui augmente les coûts en termes de capacité et de temps de calcul. De plus, les méthodes spectrales sont pour la plupart transductives, fonctionnant seulement pour le graphe donné. Il est donc impossible de déterminer la représentation de nœud inconnu d'un nouveau graphe. Ces limitations ont permis la mise en avant de nouvelles méthodes, basées sur la théorie spatiale et la transmission de messages (ou "message-passing" en anglais).

2.3.2.2 Théorie spatiale

La théorie spatiale considère le graphe comme une structure spatiale dans laquelle l'information est propagée à travers des connexions locales afin de calculer de nouvelles représentations, contrairement à la théorie spectrale qui se concentre sur la manière dont les fonctions spectrales peuvent être utilisées pour caractériser la structure du graphe. Plusieurs méthodes, notamment dérivées de certaines méthodes

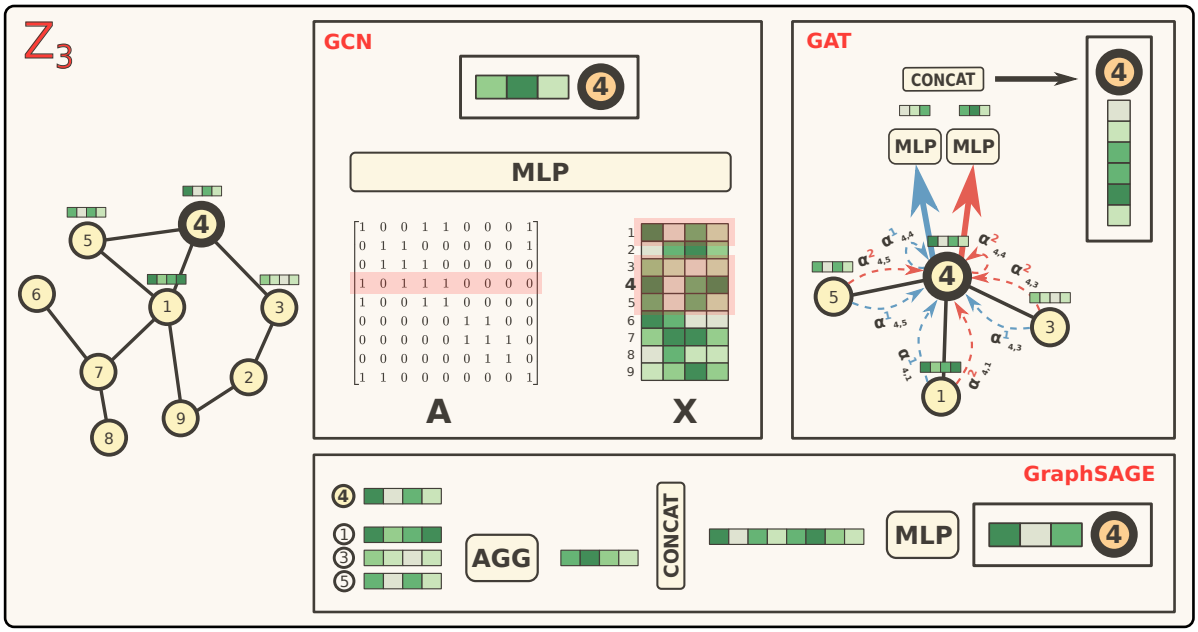

FIGURE 2.10 – *Fonctionnement des couches de convolutions des GNN. 3 approches sont explicitées, des méthodes spectrales (GCN), aux méthodes spatiales (GraphSAGE et GAT). MLP désigne un réseau de neurones de type perceptron, **AGG** une fonction quelconque d'agrégation et **CONCAT** la fonction de concaténation des nœuds. L'architecture GAT contient deux têtes d'attention, représentées chacune par les flèches rouge et bleue.*

spectrales [DBV16; KW17], peuvent être définies dans un même cadre général de transmission de messages ("message-passing") où les informations sont propagées entre les nœuds directement par les liens et de manière récursive.

La méthode GIN [Xu+19], se base sur l'isomorphisme des graphes pour différencier la structure des graphes. Elle utilise l'équation 2.2 pour définir sa convolution.

$$h_v^{(l)} = MLP^{(l)}\left(\left(1+\epsilon^{(l)}\right)h_v^{(l-1)} + \sum_{u\in\mathcal{N}_{(v)}} h_u^{(l-1)}\right) \tag{2.2}$$

L'équation 2.2, calcule la nouvelle représentation du nœud v pour la couche l. Elle commence par sommer l'ancienne représentation du nœud v $h_v^{(l-1)}$, ajustée par un paramètre d'apprentissage ϵ, avec la somme des anciennes représentations de ces voisins contenus dans $\mathcal{N}_{(v)}$. Ce nouveau vecteur est ensuite passé dans un réseau de neurones classique avec plusieurs paramètres d'apprentissage, appelé perceptron multicouche (MLP), pour finalement calculer la représentation $h_v^{(l)}$.

Le nombre de voisins par nœud n'étant pas fixe, avec un nombre de voisins pouvant être très grand, GraphSAGE [HYL17] ajoute une méthode d'échantillonnage afin d'avoir un nombre fixe de voisins pour chaque nœud. GraphSAGE se dénote aussi par sa fonction d'agrégation des nœuds voisins, présente dans sa règle de propagation (équation 2.3), et schématisée dans la figure 2.10.

$$h_v^{(l)} = \sigma\left(W^{(l)} \,.\, CONCAT\left(h_v^{(l-1)}, AGGREGATE(\{h_u^{(l-1)}, \forall u \in \mathcal{N}_{(v)}\})\right)\right) \tag{2.3}$$

La fonction d'agrégation $AGGREGATE$ agrège les caractéristiques du voisinage local d'un nœud. Elle est invariante selon les permutations des nœuds voisins. Différentes architectures peuvent être utilisées, des méthodes d'agrégation classique (moyenne) comme des réseaux de neurones récurrents, tels qu'un LSTM (adapté pour simuler l'invariance par permutation).

Velickovic et al. [Vel+18] proposent une version modifiée de cette architecture, en ajoutant une couche d'attention durant la convolution. Les mécanismes d'attention [Vas+17], utilisés dans différents domaines tels que le traitement automatique du langage (TAL), orientent le modèle, en accordant une plus grande attention à certains facteurs lors du traitement des données. Dans le cas des graphes, la couche d'auto-attention permet de donner arbitrairement plus ou moins d'importance aux nœuds voisins. Pour un nœud v, Velickovic et al. [Vel+18] commencent par apprendre les coefficients d'attention $a_{v,u} = att(\mathbf{W}h_v^{(l)}, \mathbf{W}h_u^{(l)})$ avec chaque voisin u, avec $att()$ un réseau de neurones classique à une seule couche et W la matrice de poids de cette fonction à la couche l. Ces poids sont ensuite normalisés en utilisant la fonction $SOFTMAX$, décrite par l'équation 2.4.

$$\alpha_{vu} = softmax(a_{vu}) = \frac{\exp(a_{vu})}{\sum_{w \in \tilde{\mathcal{N}}_{(v)}} \exp(e_{vw})} \tag{2.4}$$

Ces poids sont ensuite utilisés dans la fonction de propagation de la couche de convolution, comme l'équation 2.5 le montre.

$$h_v^{l+1} = \sigma\left(\sum_{u \in \tilde{\mathcal{N}}_{(v)}} \alpha_{vu} \mathbf{W} h_u^l\right) \tag{2.5}$$

GAT [Vel+18] utilise plusieurs têtes d'attention ("multi-head attention layer" en anglais) dans une même couche indépendamment, afin de stabiliser l'apprentissage et d'augmenter ces capacités. Elle concatène donc les représentations de chaque tête pour avoir la représentation finale à la couche l. L'équation 2.6 montre la fonction de propagation mise à jour en utilisant plusieurs têtes d'attention, où k représente le nombre de têtes utilisées dans la couche. Le mécanisme d'attention est schématisé dans la figure 2.10.

$$h_v^{l+1} = \|_{k=1}^{K} \sigma\left(\sum_{u \in \tilde{\mathcal{N}}_{(v)}} \alpha_{vu}^k \mathbf{W}^k h_u^{(l)}\right) \tag{2.6}$$

Plusieurs autres méthodes spatiales existent, comme la méthode GAAN [Zha+18a] qui s'inspire de l'approche GAT [Vel+18] (score d'attention supplémentaire ajouté à chaque tête d'attention utilisée), mais ne seront pas détaillées ici.

Concernant les représentations à la fin du modèle, les méthodes spatiales tendent à essayer de retenir le maximum d'informations sur les localités. Par conséquent, contrairement à l'approche originale des GCNs (équation 2.1) par exemple qui utilise seulement la dernière couche pour la représentation finale des nœuds à la fin du modèle, la plupart des méthodes spatiales [Xu+19 ; HYL17 ; Vel+18] concatènent toutes

les représentations intermédiaires des couches. Les méthodes spatiales présentent de nombreux avantages. Tout d'abord, elles sont plus modulables. L'apprentissage de la représentation d'un nœud ne nécessite pas tout le graphe, mais seulement les nœuds présents dans la localité (ou niveau de voisinage) voulue. GraphSAGE [HYL17] va plus loin en utilisant une méthode d'échantillonnage du voisinage, pour réduire encore la complexité de calcul lors de l'entrainement, ce qui la rend plus adaptée aux larges graphes. L'utilisation de lots (ou batch) est donc possible lors de l'entrainement, contrairement aux méthodes spectrales qui utilisent un unique batch comprenant tout le graphe. On note donc un gain d'espace mémoire (le graphe complet n'est pas stocké) et de temps de calcul potentiel (parallélisation de l'apprentissage). Les méthodes spatiales sont aussi dites inductives et ont donc des capacités de généralisation et de prédiction sur des graphes ou nœuds qui n'ont jamais été vus, contrairement aux méthodes spectrales [KW17] dite transductives.

2.3.2.3 Structure complexe

Les graphes peuvent aussi retenir des propriétés plus complexes, comme des poids sur les liens entre les nœuds. Les GNN peuvent prendre en compte ces poids de différentes manières. En modifiant la matrice d'adjacence d'un graphe avec la valeur du poids d'un lien, les GCN [KW17] permettent de traiter cette information. GraphSAGE [HYL17] et GAT [Vel+18] rajoutent ces poids dans la formule de propagation (équation 2.5), lorsque le lien en question est concerné. Les liens pouvant aussi avoir plusieurs caractéristiques, Gilmer et al. [Gil+17] proposent la méthode MPGNN pour modifier la fonction de propagation, en changeant le calcul de la représentation des voisins. Pour un nœud v, MPGNN utilise une fonction de transformation pour calculer la nouvelle représentation de chaque voisin u, en utilisant en entrée les caractéristiques du nœud v, celles du nœud voisin u ainsi que les caractéristiques du lien e_{vu}. Dans le cas des graphes hétérogènes (plusieurs types de liens), plusieurs solutions existent. Schlichtkrull et al. [Sch+18] proposent d'entraîner différentes matrices de poids à chaque couche, selon le nombre de types de lien. Ainsi, les représentations des nœuds voisins sont calculées en utilisant la matrice de poids du type de lien correspondant lors de la formule de propagation.

2.3.3 Représentation du graphe

La représentation des graphes (ou "graph embedding" en anglais) permet de représenter le graphe dans un sous-espace vectoriel de dimension finie, pour que ce vecteur puisse être utilisé ensuite dans des réseaux de neurones classiques, pour diverses tâches (prédiction, classification de graphe, etc.).

Il reste complexe de représenter le graphe en utilisant toutes les représentations des nœuds (section 2.3.2), dû au nombre fluctuant de nœuds (potentiellement large), entre chaque graphe. Les types de fonctions permettant de représenter un graphe dans un sous-espace peuvent être regroupées en deux catégories : les **fonctions d'agrégation** ($READOUT$) et les **fonctions de pooling**.

2.3.3.1 Fonctions d'agrégation

L'objectif des fonctions $READOUT$ est de calculer une représentation du graphe en agrégeant chaque représentation des nœuds calculés à l'étape précédente (figure 2.7).

$$h_G = READOUT\left(\left\{h_1^{(L)}, h_2^{(L)}, h_3^{(L)}, ..., h_n^{(L)}\right\}\right) \tag{2.7}$$

L'équation 2.7 illustre cette approche, où n correspond ici au nombre de nœuds, $h_1^{(L)}$ à la représentation finale du nœud v_1 après la dernière couche L (représentation concaténée de chacune des couches ou pas) et h_G à la représentation finale du graphe après application de la fonction d'agrégation $READOUT$. Des fonctions classiques [Wu+21] d'agrégation peuvent être appliquées :

1. **Somme** ou **Moyenne** : Ces fonctions, simples et peu coûteuses en temps exécution, somment (ou moyennent) les représentations de tous les nœuds. La Moyenne permet de donner une importance égale à tous les nœuds. Cependant, leur capacité à capturer des relations plus complexes entre les nœuds est limitée. Zhang et Xie [ZX20] proposent une version modifiée de GAT [Vel+18] pour représenter ces nœuds et testent ces deux fonctions pour représenter le graphe.
2. **"Max-pooling"** : Cette fonction sélectionne la représentation maximale, à partir de tous les nœuds, pour chaque caractéristique comprise dans la représentation des nœuds, ce qui facilite l'identification des nœuds les plus importants du graphe.

Des fonctions plus complexes, basées sur les réseaux de neurones, permettent de capturer des informations plus importantes du graphe.

Les réseaux de neurones récurrents (RNN), comme les LSTM, permettent d'agréger les nœuds dans un ordre séquentiel et de créer une représentation globale. Pour compenser l'absence de hiérarchie des nœuds dans un graphe, Vinyals, Bengio et Kudlur [VBK15] (Set2set) proposent d'intégrer les informations dépendantes de l'ordre dans la mémoire du LSTM. Li et al. [Li+15] (GRU-GNN) trient et sélectionnent les nœuds les plus importants, puis utilisent un autre RNN, GRU, pour concaténer les représentations en une représentation finale. Lee, Rossi et Kong [LRK18] utilisent un mécanisme d'attention sur le graphe (GAM), et calculent une pondération pour chaque nœud en fonction de son importance pour la tâche en cours, puis combinent les représentations pondérées pour donner une représentation globale.

2.3.3.2 Fonctions de pooling

Les fonctions de pooling différent légèrement. Leur but est de réduire la taille du graphe en échantillonnant le nombre de nœuds (ou de ces caractéristiques) afin de générer un graphe plus petit, jusqu'à n'avoir plus qu'une unique représentation du graphe. Zhang et al. [Zha+18b] proposent la méthode DGCNN pour trier les nœuds et sélectionner seulement les k nœuds les plus importants, afin de représenter le graphe dans une matrice de taille fixe, à la manière d'une image. Cette matrice,

ensuite, peut être passée dans un CNN afin de classifier le graphe. La fonction de tri se base néanmoins seulement sur les caractéristiques des nœuds calculés lors des convolutions, et non sur la structure en elle-même. Tixier et al. [Tix+19] considèrent aussi le graphe comme une image. Après avoir calculé la représentation de ces nœuds, ils représentent le graphe en lui appliquant une ACP, puis en créant plusieurs images en 2-dimensions à partir d'histogrammes construits sur différentes paires de dimensions de l'ACP. Ces images sont ensuite gérées sur plusieurs canaux dans un CNN-2D pour classifier le graphe. Ying et al. [Yin+18] proposent une méthode non supervisée, DiffPool, permettant de réaliser plusieurs agrégations hiérarchiques des nœuds. Cette technique permet de compresser les informations contenues dans le graphe tout en préservant la structure hiérarchique de celui-ci.

$$A^{(l+1)}\ ,\ X^{(l+1)} = POOLING\left(GNN^{(l)}_{pool}\left(A^{(l)}\right)\ ,\ GNN^{(l)}_{embed}\left(X^{(l}\right)\right) \tag{2.8}$$

DiffPool [Yin+18] applique une couche de pooling différentiable sur les nœuds du graphe, comme le montre l'équation 2.8. Cette couche de pooling est composée de deux parties : la première calcule, à partir d'un GNN à convolution ($GNN^{(l)}_{pool}$), les nouvelles caractéristiques de chaque nœud du graphe en entrée de la couche l. La seconde partie ($GNN^{(l)}_{embed}$) permet de regrouper les nœuds afin de former des clusters de nœuds similaires. Une fois les scores de chaque nœud calculés et regroupés en clusters, la fonction $POOLING$ agrège, pour chaque cluster, la représentation de ces nœuds afin de n'avoir qu'une seule représentation pour chaque cluster. Ces clusters deviennent ensuite des nœuds dans le nouveau graphe de sortie. Ce processus est répété de manière récursive jusqu'à ce qu'on atteigne le niveau de granularité désiré, généralement jusqu'à ce qu'on atteigne un graphe contenant un seul nœud, qui correspondra à la représentation finale du graphe. Cette opération de $POOLING$ résulte en un graphe plus petit, mais préservant la structure hiérarchique du graphe d'origine, et permet de réduire la complexité des réseaux de graphes et de mieux gérer les données de grande dimension. Toutes ces méthodes permettent donc de représenter le graphe en se basant sur les informations pertinentes extraites par les GNN pour ces nœuds. Néanmoins, l'inconvénient majeur réside dans sa représentation réduite du graphe en un vecteur, qui représente une grosse perte d'informations, pouvant amener à une mauvaise représentation du graphe. La figure 2.11 regroupe les différentes approches autour de la représentation des nœuds et du graphe à l'aide des réseaux de neurones graphiques présentés lors de cet état de l'art.

Dans les sections 2.3.2 et 2.3.3, nous avons présenté différentes méthodes GNN afin de représenter les nœuds d'un graphe, ainsi que le graphe lui-même, sous la forme d'un vecteur dans un espace euclidien. Ces méthodes peuvent être entrainées selon différentes tâches d'apprentissage, ce que nous présentons dans la section 2.3.4.

2.3.4 Tâches d'apprentissage

Les tâches d'apprentissage effectuées par les réseaux de neurones graphiques diffèrent peu des réseaux de neurones classiques de type MLP. Les fonctions de pertes

REPRESENTATION DES NOEUDS		REPRESENTATION DU GRAPHE	
Théorie spectrale	*Théorie spatiale*	*READOUT*	*Pooling*
Spectral CNN [Bru+14]	GIN [Xu+19]	GAT-GC [ZX20]	DGCNN [Zha+18b]
ChebNet [DBV16]	GraphSAGE [HYL17]	SET2SET [VBK15]	GC-2D [Tix+19]
GCN [KW17]	GAT [Vel+18]	GRU-GNN [Li+15]	Diffpool [Yin+18]
DGCN [ZM18]	GAAN [Zha+18a]		
VGAE [KW16]	MPGNN [Gil+17]		
	RGCN [Sch+18]		
	DeepWalk [PAS14]		

FIGURE 2.11 – *Comparaisons de différentes approches autour des GNN.*

(ou "loss function" en anglais) sont adaptées aux caractéristiques des graphes. La plupart des travaux se basent sur des fonctions répandues dans l'apprentissage des réseaux de neurones [Wu+21].

2.3.4.1 Apprentissage supervisé et semi-supervisé

Les méthodes d'apprentissage supervisées permettent d'entrainer un modèle à partir d'exemples déjà labellisés. Dans le cas des graphes, elles peuvent concerner des tâches de classification ou de prédiction autour de trois objets, illustrées dans la figure 2.7 : les nœuds, les liens et les graphes. Concernant la classification des nœuds, chaque nœud contient donc son label. La plupart des méthodes [DBV16; Vel+18; KW17] minimisent une fonction de "cross-entropy" adaptée aux graphes, lors de la phase d'entrainement. Cette fonction, présentée dans l'équation 2.9, mesure la divergence entre la distribution de probabilité prédite et la distribution de probabilité réelle d'un modèle de classification.

$$\mathcal{L}_{CE} = -\frac{1}{n}\sum_{v=1}^{n}\sum_{c=1}^{C} y_{vc}\log\left(\hat{y}_{vc}\right) \tag{2.9}$$

$\mathcal{L}_{CE}$ représente ici la perte de "cross-entropy", n est le nombre de nœuds dans le graphe, C correspond au nombre de classes, y_{vc} est l'étiquette (0 ou 1) pour le nœud v et la classe c, et $\hat{y}_{vc}$ est la prédiction pour le nœud v et la classe c. Dans le cas de l'apprentissage semi-supervisé, seul l'apprentissage diffère, car seulement une partie des nœuds est labellisée. Les méthodes spectrales, comme les GCNs [KW17], étant transductives et ne pouvant être appliquées qu'à un seul graphe, sont principalement concernées.

Dans le cas de la classification des liens, la différence principale repose sur la manière de représenter le lien. Les nouvelles représentations des deux nœuds concernés par le lien sont calculées par un GNN, puis sont concaténées. Si le lien a des caractéristiques supplémentaires, ces caractéristiques sont aussi concaténées à la suite. Après quoi la représentation du lien est passée dans un réseau de neurones classique. Finalement, la

fonction de perte est calculée puis minimisée lors de l'entraînement à partir des labels des liens.

Dans le cas de la classification de graphes, le jeu de données est différent. Il contient plusieurs graphes différents et labellisés. Cependant, après avoir calculé la représentation du graphe (figure 2.7), comme indiqué dans la section 2.3.3, l'apprentissage demeure le même que pour la classification des nœuds, à la différence que les fonctions de pertes sont calculées sur les prédictions émises sur les graphes et non les nœuds.

2.3.4.2 Apprentissage auto-supervisé

L'apprentissage auto-supervisé permet à un modèle d'être entraîné sans avoir de données labellisées, en générant des labels à partir d'informations propres aux données. Dans le cas des graphes, on cherche à utiliser les relations entre les nœuds d'un graphe pour capturer des caractéristiques intrinsèques au graphe [Xie+22]. La prédiction des liens est l'une des méthodes les plus répandues. Elle consiste à prédire, à partir d'un graphe, si un lien entre deux nœuds existe. Une première approche utilise des paires de liens lors de la phase d'apprentissage, générant deux jeux de données : le premier contient des paires de nœuds existantes (lien présent dans le graphe), et le second contient des paires non-existantes. La représentation de chaque paire (lien) est générée après avoir calculé les nouvelles représentations à partir d'un modèle GNN pour chaque nœud. Plusieurs approches existent pour représenter ce lien :

- Des fonctions simples, comme le calcul du produit scalaire entre les deux nœuds, afin d'obtenir un score qui une fois normalisé servira de prédiction ;
- Des fonctions plus complexes, comme l'application d'un réseau de neurone MLP, qui utilise comme entrée un vecteur concaténant la représentation des deux nœuds, et qui prédira finalement l'existence ou non du lien.

D'autres approches autour de la prédiction de liens existent [Xie+22], tels que celle basée sur la fonction de coût par triplet. Cette méthode consiste à entraîner un modèle en lui présentant des triplets de nœuds. Le modèle doit apprendre à prédire si deux nœuds sont connectés par une arête en maximisant la marge entre la similarité de la paire de nœuds connectés et la similarité de la paire de nœuds non connectés. L'objectif est de renforcer les liens entre les nœuds similaires et de diminuer ceux des nœuds dissemblables. Le principal problème de ces approches réside dans le déséquilibre des classes. Le nombre de liens non-existants étant généralement bien supérieur à ceux existant, le risque de sur-apprentissage est présent. Des méthodes d'échantillonnage ("sampling") permettent en partie de réduire ce biais.

2.3.4.3 Apprentissage non supervisé

L'apprentissage non supervisé permet à un modèle d'être entraîné seulement à partir de schémas dans les données, sans la nécessité d'avoir des données labellisées au préalable. Perozzi, Al-Rfou et Skiena [PAS14] proposent une première approche non supervisée, basée sur les marches aléatoires ("Deepwalk"), afin de représenter les nœuds en fonction de leur similarité (proximité). Elle consiste à créer plusieurs marches aléatoires, de taille fixe, dans le graphe, partant de chaque nœud $v \in V$.

Ensuite, à la manière d'un Word2vec [Mik+13] pour le texte, un réseau de neurones de type MLP apprend à prédire le voisinage de chaque nœud. La couche intermédiaire de ce modèle sert ensuite à représenter les nœuds.

KIPF et WELLING [KW16] proposent un modèle de type auto-encodeur VGAE, pour représenter les nœuds (et graphe) dans un sous-espace vectoriel. Un réseau de neurones de type GNN encode les nœuds dans des représentations latentes de faible dimension, puis encode le graphe dans une représentation latente de faible dimension. La représentation latente du graphe est ensuite décodée pour reconstruire le graphe original. Le processus d'encodage et de décodage est optimisé à l'aide d'une fonction de perte qui mesure la différence entre le graphe original et la reconstruction. Les différentes représentations latentes serviront à représenter nos nœuds ainsi que le graphe.

Dans la section 2.3.4, nous avons étudié différentes tâches d'apprentissages autour des GNN, avec des méthodes supervisées, auto-supervisées et non-supervisées. Dans ce manuscrit, ces approches seront étudiées dans le contexte de l'analyse de la controverse sur les médias sociaux dans les chapitres 5 et 4.

2.4 Présentation des différents types et jeux de données

Dans cette section, nous présenterons les différents types de données étudiées dans la littérature, avec une présentation approfondie des jeux de données utilisés dans nos travaux.

Concernant l'analyse des pages web (section 2.2.1), plusieurs types de données existent. Les données Wikipédia sont les plus utilisées dans ces études, avec des jeux de données comprenant les articles eux-mêmes, leur titre, ainsi que les métadonnées de leurs révisions [Kit+07 ; Vuo+08]. Les jeux de données autour des médias en ligne (section 2.2.1.2) sont aussi analysés. SRITEJA, PANDEY et PUDI [SPP17] utilisent différents articles, provenant de médias principaux (ABC News, CNN, Politico), ou de médias plus subjectifs ("Addicting Info", "Right Wing News"). Un jeu de données comprenant des articles du journal britannique "The Guardian" est aussi utilisé dans l'analyse des médias en ligne [KA19].

Concernant les médias sociaux, qui constituent le cœur de nos travaux, plusieurs jeux de données ont été présentés dans la littérature. Nous porterons un intérêt particulier sur trois jeux d'entre eux (provenant de Reddit et Twitter), que nous utiliserons dans nos travaux.

2.4.1 Ensemble de données 1

Ce premier jeu de données correspond à des données réelles, provenant du réseau social Reddit. Reddit est un forum social américain. Les utilisateurs peuvent soumettre des contenus sur le site, tels que des liens, des messages textuels, des images et des

Table 2.2 – *Statistiques sur le contenu des 6 subreddits provenant de Reddit.*

	AM	AW	FN	LS	PF	RS
Nombre de posts	3305	2969	3934	1573	1004	2248
Nombre moyen d'utilisateurs par post	72	67	76	79	47	48
Nombre moyen de commentaires par post	144	141	159	132	95	98
Nombre moyen de mots par commentaire	41	42	34	28	52	61
% de commentaires ayant un nombre de tokens ≥ 256	2,68	2,64	1,61	1,03	4,1	6,17

vidéos, qui sont ensuite approuvés ou rejetés et commentés par les autres membres. Les posts sont organisés par thème, appelés subreddits, qui couvrent une variété de sujets tels que l'actualité, la politique, la religion, la science, les films, etc.

Il a été collecté lors des travaux réalisés par Hessel et Lee [HL19], puis réutilisé dans les travaux couverts par Zhong et al. [Zho+20]. Les données collectées comportent seulement du contenu en anglais, et couvrent une période allant de 2007 à février 2014. L'ensemble de données est divisé en 6 jeux de données distincts, correspondant à six subreddits spécifiques : *AskMen* (AM), *AskWomen* (AW), *Fitness* (FN), *LifeProTips* (LT), *PersonnalFinance* (PF) et *Relationships* (RS).

Sur Reddit, chaque utilisateur peut commenter un post lié à un subreddit spécifique. Chaque subreddit contient un ensemble de posts. Chacun de ces posts est représenté par le contenu du post lui-même, le fil de discussion (ou arbre de commentaire) associé contenant les commentaires des autres utilisateurs à propos de ce même post, ainsi que certaines métadonnées sur chacun de ces messages (du post et des commentaires). Enfin, seuls les posts comportant un total d'au moins 30 commentaires sont conservés, en partant du principe que le fait d'avoir moins de 30 commentaires n'est pas une condition suffisante pour construire un graphe significatif. Chaque post est automatiquement étiqueté comme controversé ou non controversé, en fonction de diverses métadonnées sur le post [HL19], notamment le ratio entre les votes positifs et négatifs. En effet, les posts ayant reçu de nombreux votes positifs et négatifs devraient être considérés comme les plus controversés.

Le tableau 2.2 présente des statistiques sur les six subreddits de cet ensemble de données.

2.4.2 Ensemble de données 2

Ce second jeu de données correspond à des données réelles, provenant du réseau social Twitter. Twitter est un réseau social et de micro-blogging, où les utilisateurs publient des tweets et interagissent avec eux. Les utilisateurs peuvent publier, aimer, retweeter et citer différents tweets. Twitter présente de nombreux avantages par rapport aux autres médias sociaux. Tout d'abord, l'action du Retweet, qui consiste à repartager le contenu exact d'un tweet, est considérée comme une approbation. Cela signifie qu'un utilisateur est d'accord avec le contenu du tweet original et qu'il le soutient. De plus, Twitter est l'un des médias sociaux les plus utilisés pour les débats publics et est également utilisé pour rapporter des informations sur des événements d'actualité. Par conséquent, les données disponibles sont plus importantes que celles de n'importe quel

TABLE 2.3 – *Informations sur les données mises à disposition par* GARIMELLA *et al.* [Gar+18a] *et l'API de Twitter, pour chaque sujet séparément. Les neuf premiers sujets représentent des sujets controversés, tandis que les six derniers représentent des sujets non controversés. Le ratio de récupération correspond au ratio des tweets récupérés depuis l'API comparés à tous les identifiants existant.*

Sujet	Tweets récupérés	Ratio de récupération (%)	# principal	Description (2015)
LEADERSDEBATE	693 281	60,9%	#leadersdebate	Débat lors des élections nationales au Royaume-Uni, 3 Mai
RUSSIA_MARCH	51 395	43,3%	#russia_march	Manifestations après la mort de Boris Nemtsov, 1-2 mars
UKRAINE	155 258	54,0%	#ukraine	Conflit en Ukraine, 27 février-2 mars
INDIANA	68 008	58,4%	#indiana	Une pizzeria de l'Indiana refuse d'organiser un mariage gay, 2-5 avril
BEEFBAN	46 654	55,2%	#beefban	Interdiction de la viande bovine par le gouvernement indien, 2-5 mars
NETANYAHU	113 343	44,5%	#netanyahuspeech	Discours de M. Netanyahu devant le Congrès américain, 3-5 mars
INDIASDAUGHTER	92 289	55,0%	#indiasdaughter	Documentaire indien controversé, 1-5 mars
BALTIMORE	90 908	41,7%	#baltimoreriots	Émeutes à Baltimore après la mort d'un Noir par la police, 28-30 avril
NEMTSOV	105 379	57,4%	#nemtsov	Décès de Boris Nemtsov, 28 février-2 mars
ONEDIRECTION	190 662	38,0%	#1dfamheretostay	Concert des OneDirection, 27-29 mars
SXSW	233 810	68,0%	#sxsw	Conférence SXSW, 13-22 mars
MOTHERSDAY	1 033 839	57,5%	#mothersday	Fête des mères, 8 mai
GERMANWINGS	561 836	61,9%	#germanwings	Crash du vol Germanwings, 24-26 mars
NEPAL	772 618	59,5%	#nepal	Tremblement de terre au Népal, 26-29 avril
ULTRALIVE	187 920	51,6%	#ultralive	Ultra Music Festival, 18-20 mars

autre réseau social. Cela nous permet de travailler avec une quantité d'informations non négligeable et plus représentative, réduisant ainsi la variance de nos données. Cependant, il reste compliqué de capturer uniquement les tweets d'un sujet en particulier à partir de mots-clés. En effet, des tweets contenant ces mots-clés peuvent alors appartenir au jeu de données sans nécessairement appartenir au sujet concerné.

Nous avons récupéré nos données à partir des indications faites par GARIMELLA et al. [Gar+18a]. En effet, les auteurs ont sélectionné dans leurs travaux 20 sujets différents (10 controversés et 10 non controversés), liés à différents événements ou personnes publiques. Ces sujets ont été labellisés controversés ou non à partir d'une analyse faite par les auteurs sur les médias et journaux en ligne. Après avoir sélectionné les hashtags communs liés à chaque sujet, les auteurs ont réalisé des requêtes de recherche, à partir de ces mots-clés, sur l'API de Twitter [7]. Les tweets collectés par les auteurs ont été publiés entre le 27 février et le 15 juin 2015. Cependant, seuls les identifiants des tweets de 15 sujets (dont 9 controversés) ont été mis à disposition [8] par GARIMELLA et al. [Gar+18a]. Nous avons procédé à la collecte des données à partir des identifiants des tweets sur l'API de Twitter. Or certains tweets n'étaient plus accessibles. Différentes raisons peuvent expliquer cela, notamment le fait qu'ils puissent avoir été supprimés par leurs auteurs respectifs ou par des modérateurs de contenu.

Après avoir collecté les tweets encore disponibles ainsi que leur méta-données, Nous avons obtenu un taux global de 56,5% de tweets retrouvés, avec une moyenne de 293 146 tweets par sujet. Cela a été considéré comme suffisant pour la fiabilité de nos expérimentations. Différentes statistiques sur ce jeu de données sont présentées dans le tableau 2.3.

7. https ://developer.twitter.com/en/docs/twitter-api
8. L'ensemble de ces identifiants Twitter est disponible sur le dépôt https://github.com/gvrkiran/controversy-detection

Table 2.4 – *Statistiques sur l'ensemble des données extraites par sujet. Les 15 premiers sujets représentent des sujets controversés, tandis que les 15 derniers représentent des sujets non controversés.*

Sujet	Période	# Tweets	# Utilisateurs	# Retweets	Mot-clé (Description) [LANG]
impeachment-5-10	31oct–10nov, 2015	123 697	20 878	51 404	Roussef impeachment
menciones-1-10enero	1–11jan, 2018	81 209	25 591	49 034	Macri's mentions
menciones-11-18marzo	11–18mar, 2018	406 869	31 659	58 797	Macri's mentions
menciones-20-27marzo	24–26mar, 2018	97 950	34 975	68 990	Macri's mentions
menciones-05-11abril	5–10apr, 2018	220 460	63 358	144 600	Macri's mentions
menciones05-11mayo	5–10may, 2018	267 283	63 030	146 217	Macri's mentions
bolsonaro27	27oct, 2018	120 162	45 629	88 160	Brazilian elections
bolsonaro28	28oct, 2018	151 952	84 986	104 955	Brazilian elections
bolsonaro30	30oct, 2018	174 565	73 399	130 599	Brazilian elections
kavanaugh06-08	8oct, 2018	157 721	71 933	123 055	Kavanaugh's nomination
kavanaugh16	3oct, 2018	168 571	66 765	131 270	Kavanaugh's nomination
kavanaugh02-05	5oct, 2018	181 202	74 834	145 476	Kavanaugh's nomination
lula_moro_chats	10–11jun, 2019	199 423	66 462	143 318	Lula's mentions during Moro chats news
leadersdebate	11–21nov, 2019	250 000	76 863	174 466	Candidates debate
pelosi	6dec, 2019	252 000	95 558	209 044	Trump Impeachment
area51	3–13jul, 2019	178 220	107 460	156 481	Jokes about Area51
otdirecto20e	13–20jan, 2020	148 061	25 436	95 321	Event of a Music TV program in Spain
VanduMuruganajith	23jun, 2019	167 434	8 401	113 208	Ajith's fans
nintendo	19–28may, 2019	166 145	94 255	105 793	Nintendo's release
messicumple	23–24jun, 2019	177 770	98 448	128 099	Messi's birthday
wrestlemania	8apr, 2019	213 355	61 051	106 347	Wrestlemania event
kingjacksonday	24–27mar, 2019	142 240	39 838	107 298	popstar's birthday
notredam	16apr, 2019	171 306	99 346	146 280	Notredam fire
Thanksgiving	28nov, 2019	250 000	155 358	164 174	Thanksgiving day
halsey	7–8jun, 2019	237 501	98 008	204 149	Halsey's concert
feliznatal	25–26dec, 2019	305 879	193 989	212 893	Happy Christmas wishes
exodeux	7nov, 2019	179 908	37 384	135 579	EXO's new album
bigil	21–22jun, 2019	205 557	25 830	171 322	Vijay's birthday
championsasia	24nov–1dec, 2019	221 925	68 754	145 829	Al-Hilal champion
seungwoobirthday	23dec, 2018	251 974	18 977	193 183	Segun Woo singer birthday

2.4.3 Ensemble de données 3

Ce troisième jeu de données est aussi basé sur des données réelles, provenant du réseau social Twitter. Ce jeu contient des tweets et retweets sur un total de 30 sujets différents fournis par Zarate et al. [Zar+20], récupérés à l'aide de l'API Twitter, entre 2019 et 2020.

Quinze sujets ont été manuellement étiquetés comme controversés et quinze comme non controversés, à partir de multiples sources provenant de différents médias en ligne [Zar+20]. Les sujets non controversés se concentrent sur des domaines tels que le divertissement ou les événements notables sans controverse (scientifique, sportif, etc.). À l'inverse, les sujets controversés sont principalement axés sur des événements politiques (élections, affaires judiciaires), où la controverse et la polarisation des communautés est souvent présente. Ce jeu rassemble des tweets dans six langages différents, représentant cinq régions du monde, permettant ainsi d'étudier la controverse en limitant l'impact de l'environnement des utilisateurs.

Chaque sujet contient des tweets (et des retweets) récupérés à partir de hashtags (ou mots-clés), correspondant à l'événement en question. Plusieurs informations sont

extraites des tweets, tels que l'identifiant de l'utilisateur, le texte et les informations sur l'utilisateur du retweet, s'il s'agit d'un retweet. Seuls les tweets originaux retweetés au moins une fois sont conservés, ainsi que les utilisateurs concernés. Il est à noter que la plupart des utilisateurs ne font que retweeter et ne publient jamais de tweets originaux. D'après les données auxquelles nous avons eu accès, certains tweets peuvent manquer dans notre ensemble de données, en fonction du sujet, car les tweets peuvent avoir été supprimés, comme expliqué dans la section 2.4.2. Ce jeu se compose donc de 30 sujets dont le nombre de tweets est compris entre 5 458 et 36 716, avec un nombre d'utilisateurs compris entre 3 696 et 161 612 par sujet. Les détails sur ces 30 sujets sont présentés dans la table 2.4.

2.5 Conclusions

Dans la section 2.2, nous avons présenté différents travaux autour de la détection, de la quantification et de l'explication de la controverse, que ce soit dans des médias et articles en ligne (section 2.2.1), ou dans les médias sociaux (section 2.2.2) où les utilisateurs interagissent et participent à des discussions. Nous avons résumé cette profusion de travaux et leur évolution progressive dans un tableau récapitulatif.

Ensuite, dans la section 2.3, nous avons présenté les concepts et théories derrière les GNN et montré la multitude d'architectures disponibles en apprentissage profond et plus spécialement via des réseaux de neurones graphiques. Pour ce faire, nous avons produit différentes figures originales pour les comparer, mais aussi afin de comprendre l'apport de ces architectures pour les données non structurées. Les sections 2.3.2 et 2.3.3 présentent différentes approches et modèles afin de pouvoir représenter les nœuds et les graphes, alors que la section 2.3.4 se concentre sur les différentes tâches d'apprentissage de ces modèles.

Enfin, la section 2.4 décrit des jeux de données provenant de la littérature, avec notamment une présentation approfondie de trois jeux de données utilisés lors de nos différentes expérimentations.

Nos travaux présentés dans les chapitres 4 et 5 combinent l'utilisation de ces réseaux de neurones graphiques sur des tâches de détection et quantification de la controverse. Le chapitre 3 suivant traite lui de l'explication des sujets controversés par l'analyse du contenu textuel.

CHAPITRE 3

EXPLICATION DE LA CONTROVERSE PAR L'ANALYSE DES COMMUNAUTÉS POLARISÉES SUR TWITTER

Sommaire

3.1 Introduction

Dans ce chapitre, les travaux présentés contribuent principalement à l'explication de la controverse. Nous analysons les discussions sur Twitter du point de vue des communautés d'utilisateurs, afin d'étudier l'impact du texte dans la classification des tweets dans la communauté de leur utilisateur respectif. Nous proposons un pipeline, basé sur les valeurs de SHAP, pour quantifier la controverse sur Twitter. Nous analysons les contributions des caractéristiques textuelles ayant un impact sur les prédictions de trois classifieurs de tweets. Les résultats que nous obtenons à partir des diagrammes SHAP et de l'analyse statistique montrent clairement l'impact significatif de certaines caractéristiques textuelles dans la classification des tweets. Ils mettent également en évidence la pertinence de l'étude ainsi que les avantages potentiels de la combinaison du texte et des interactions avec les utilisateurs pour quantifier la controverse. Ces travaux ont mené à une publication dans la conférence IDEAS, paru en 2023 [Ben+23c].

Le reste du chapitre est organisé comme suit. La section 3.2 introduit le contexte et les motivations de l'approche, ainsi que les contributions réalisées. La section 3.3 présente la méthodologie que nous proposons pour l'analyse des sujets controversés et de leurs communautés. La section 3.4 présente la préparation des données et introduit en détail les deux sujets traités en profondeur. La section 3.5 regroupe les expériences et les résultats obtenus. Enfin, la section 3.6 recense les travaux effectués et conclut le chapitre.

3.2 Contexte

La controverse peut être étudiée selon différentes perspectives. Dans ce chapitre, la controverse fait référence à un sujet attirant différents points de vue, ainsi qu'à des réactions positives et négatives sur un événement spécifique, scindant les utilisateurs en différentes communautés. La recherche sur la controverse a donné lieu à deux grandes catégories de travaux (figure 2.6) : la détection/quantification de la controverse et l'explication de la controverse. Alors que la première vise à quantifier la controverse sur un sujet, la seconde vise à comprendre pourquoi un sujet est controversé. Dans ce chapitre, nous analysons la controverse pour des données Twitter.

En apprentissage profond, il est impossible pour un être humain de saisir toutes les nuances des décisions issues des réseaux, comprenant souvent des millions de neurones. Le but est d'analyser les résultats des prédictions de nos modèles, et d'identifier les caractéristiques qui poussent le modèle à produire ces prédictions. Que la prédiction soit fausse ou non, les contributions des caractéristiques sur la prédiction permettent de mieux comprendre les modèles, mais aussi les tâches qu'on essaye d'accomplir. Plusieurs approches, plus ou moins complexes, étudient cette explicabilité des réseaux de neurones. L'analyse de sensibilité ("sensibility analysis" en anglais) crée par exemple un score de pertinence pour chaque caractéristique d'entrée du modèle, à partir des dérivées partielles des représentations de celles-ci et de la rétropropagation des gradients [RHW+85].

Dans ce chapitre, nous examinons l'explication de la controverse en utilisant la

méthode SHAP [LL17] pour mesurer équitablement la contribution de chaque caractéristique textuelle des tweets à la détection des communautés controversées. SHAP, considérée comme une contribution importante dans le domaine de l'explicabilité des intelligences artificielles, est utilisée afin de comprendre comment un modèle donné génère ses prédictions. Il s'agit d'une méthode agnostique, ce qui signifie qu'elle peut être utilisée pour expliquer les prédictions et les classifications de n'importe quel modèle d'apprentissage automatique. SHAP est fondée théoriquement et exploite le concept des valeurs de Shapley et de la théorie des jeux coopératifs [Sha52]. Le concept de valeur de Shapley est un moyen de répartir équitablement la récompense d'un jeu entre les joueurs qui contribuent au résultat du jeu. Le terme "équitablement" est défini mathématiquement, ce qui signifie que la fonction de redistribution de la récompense satisfait quatre propriétés :

- **Efficacité** : Garantit une distribution complète du résultat entre les caractéristiques ;
- **Symétrie** : Garantit que deux caractéristiques contribuant de manière égale ont la même récompense ;
- **Nullité** : Garantit des récompenses nulles pour les caractéristiques qui ne contribuent pas au résultat ;
- **Additivité** : Considère la récompense additive d'une caractéristique en présence de plusieurs résultats du jeu.

À notre connaissance, notre étude est la première tentative utilisant SHAP pour des besoins d'explication de la controverse dans les médias sociaux. SHAP nous servira à trouver les caractéristiques textuelles contribuant le plus à la prédiction de l'appartenance d'un tweet à une communauté.

Dans nos travaux, nous allons également utiliser une ressource pour saisir la relation entre l'utilisation des mots dans un texte et les états cognitifs et mentaux de l'auteur du texte. LIWC [Boy+22] (de l'anglais "Linguistic Inquiry and Word Count") a été mis au point par des psycholinguistes. L'outil analyse des milliers de caractéristiques afin d'obtenir un score pour chacune des dimensions. De nombreux travaux de recherche ont été réalisés en utilisant les caractéristiques de LIWC, tels que les travaux sur l'identification et l'analyse automatique d'affirmations [Nak+22], ou de plaintes [PGA19]. À notre connaissance, [KWH21] est le seul travail qui prend en compte l'analyse des caractéristiques du texte pour la détection des sujets controversés. Les caractéristiques des discussions (usage des mots, style d'écriture) ont été étudiées pour mesurer le pouvoir prédictif de ces caractéristiques sur la controverse au sein des fils de discussions Reddit.

Contributions. Nous nous intéressons à l'étude de la controverse dans les discussions sur Twitter. La subjectivité d'un tel concept étant problématique, nous adoptons un point de vue plus large en appliquant une analyse du point de vue des communautés. Notre contribution est donc double :

1. **Quantification de la controverse**. Nous commençons par **quantifier** la controverse sur les discussions (sujets), en utilisant à la fois les propriétés structurelles

et textuelles des tweets, en montrant que les informations textuelles contiennent des caractéristiques intéressantes pour aider à quantifier la controverse.

2. **Explication de la controverse**. Nous proposons ensuite une solution pour **expliquer les sujets controversés**, à travers leurs communautés, en étudiant la contribution des caractéristiques sur différents modèles de classification des tweets dans la communauté correspondante à l'auteur original du tweet. Cette analyse est faite en utilisant SHAP, à la fois sur des caractéristiques textuelles et conceptuelles (LIWC). Nous étudions cette solution en l'appliquant à deux sujets et montrons que l'analyse génère des résultats prometteurs généralisables à différents sujets.

3.3 Méthode

Comme indiqué précédemment, nous allons explorer la controverse du point de vue des communautés d'utilisateurs concernant un sujet, en analysant le texte contenu dans les tweets. Afin d'expliquer et de comprendre les communautés autour de ces sujets controversés, nous proposons un pipeline composé de quatre étapes, comme indiqué dans la figure 3.1.

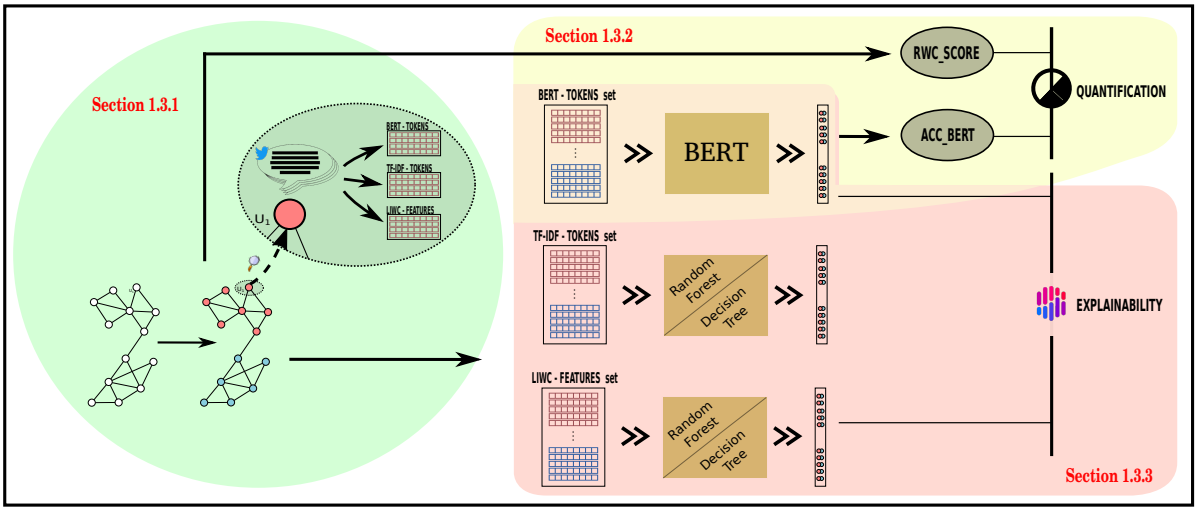

FIGURE 3.1 – *Présentation du pipeline utilisé pour quantifier et expliquer la controverse à partir de l'analyse de la communauté.*

La section 3.3.1 décrit la manière dont le graphe est construit, ainsi que les communautés. Elle décrit aussi la manière dont les tweets sont représentés pour chaque utilisateur. La section 3.3.2 détaille la manière dont la controverse d'un sujet est quantifiée, à l'aide de métriques basées à la fois sur les données structurelles et textuelles. Enfin, la section 3.3.3 décrit comment les sujets controversés sont expliqués, à partir de l'analyse des communautés d'utilisateurs, et l'analyse SHAP des contributions des caractéristiques textuelles aux prédictions des modèles.

3.3.1 Création du graphe utilisateur

Construction du graphe et partitionnement en communautés. Considérons qu'un sujet T est représenté par un ensemble de tweets (y compris les retweets) $T = t_1, t_2, \ldots, t_r$. t_i désigne le i^{eme} tweet du sujet T_j. Chaque sujet T_j est représenté sous la forme d'un graphe non orienté de retweets d'utilisateurs, dans lequel les nœuds représentent les utilisateurs, et deux utilisateurs (nœuds) sont connectés si l'un d'eux a retweeté l'autre. Pour garantir la fiabilité du partitionnement, seule la plus grande composante connectée est conservée dans le graphe final. En effet, certains groupes d'utilisateurs peuvent être déconnectés des autres.

Afin d'étiqueter les utilisateurs en fonction de leurs communautés respectives pour chaque sujet, nous nous appuyons sur les travaux de Garimella et al. [Gar+18a], et utilisons l'algorithme de partitionnement metis [KK95] pour partitionner chaque graphe en deux communautés. Dans cette étude, nous ne considérons que les deux plus grandes communautés. Elles correspondent aux soutiens et opposants à un sujet controversé. Les sous-communautés ne sont pas prises en compte. Chaque utilisateur est étiqueté par le label de la communauté (C_0 ou C_1) à laquelle il appartient.

Traitement du contenu des tweets. Les utilisateurs rassemblés dans le graphe peuvent être les auteurs d'un ou de plusieurs tweets, comme d'aucun s'ils ne publient que des retweets. Chaque tweet est étiqueté avec l'étiquette de son auteur original (correspondant à sa communauté, C_0 ou C_1). Les tweets des utilisateurs qui ne sont pas connectés dans le graphe final sont écartés de notre analyse.

Pour chaque tweet original, différents types de caractéristiques peuvent être créés à partir de leur contenu. Dans notre approche, nous avons considéré trois types de caractéristiques, générées par les méthodes TF-IDF, BERT et LIWC, mais tout autre type de caractéristiques peut être ajouté. Enfin, trois ensembles de caractéristiques sont générés, "BERT-TOKENS", "TF-IDF-TOKENS" et "LIWC-FEATURES". Les trois types de caractéristiques utilisés sont présentés ci-dessous.

3.3.1.1 Caractéristiques textuelles

Les deux premiers ensembles de caractéristiques sont créés à partir du contenu textuel des tweets.
TF-IDF-TOKENS. TF-IDF (de l'anglais Term Frequency-Inverse Document Frequency) est une méthode statistique permettant de mesurer l'importance d'un mot dans un document par rapport à un corpus de texte, en pondérant la fréquence du terme dans le document par rapport à sa rareté dans le corpus. Elle est couramment utilisée dans la recherche d'informations et l'exploration de textes pour l'extraction de caractéristiques et la classification de textes. Nous avons établi le dictionnaire du corpus à partir du même ensemble de données d'entrainement que celui utilisé pour la classification du modèle, décrit dans les sections 3.3.2 et 3.3.3. Chaque tweet est encodé indépendamment des autres.
BERT-TOKENS. En se basant sur le générateur de jetons ("Tokenizer" en anglais) du modèle BERT, chaque tweet est traité, puis représenté par un ensemble de tokens. Le générateur de jetons de BERT est une étape de prétraitement dans les modèles de type BERT [Dev+19], qui tokenise le texte d'entrée en faisant correspondre chaque mot à

TABLE 3.1 – *Description des deux premiers niveaux des caractéristiques de LIWC.*

1^{er} **niveau**	2^{ieme} **niveau**
SUMMARY DIMENSION	WC (word count), WPS (word per sentence), BigWords, Dictionary word count, Analytic, Clout, Authentic, Tone
LINGUISTIC	Function (pronoun, determinant, adverb...), Verb, Adj, Quantity
PUNCTUATION MARKS	period, Comma, QMark, exclam, Apostro, OtherP
PSYCHOLOGICAL PROCESSES	Drives (affiliation, power), Cognition, Affect (emotion), Social (behavior & references)
EXPANDED DICTIONARY	Culture, Lifestyle, Physical, States, Motive, Perception, Time orientation, Conversation

un index unique, en ajoutant des tokens spéciaux pour séparer les phrases et en codant le texte à l'aide de sous-mots pour les mots hors vocabulaire. Il permet de représenter le texte passé dans le modèle BERT présenté dans la section 3.3.2.

3.3.1.2 Caractéristiques conceptuelles

Comme indiqué précédemment, notre objectif est de trouver des caractéristiques significatives pouvant être expliquées, et qui peuvent aider à comprendre la controverse dans les médias sociaux.
LIWC-FEATURES. LIWC [Boy+22] est un outil qui analyse le contenu textuel en aidant à comprendre différents états psychologiques tels que les pensées, les sentiments ou la personnalité, permettant d'obtenir de nouvelles informations, sur la base de l'étude statistique de l'utilisation des mots. Ces caractéristiques sont organisées de manière hiérarchique et classées par catégorie. Les deux premiers niveaux de caractéristiques analysées par LIWC [1] sont décrits dans le tableau 3.1.

Chaque caractéristique possède son propre dictionnaire reflétant une catégorie psychologique. Les scores renvoyés sont calculés à partir de la proportion de mots dans leur dictionnaire respectif par rapport à ceux du contenu, et varient entre 0 et 100 (normalisés par le nombre total de mots). Seules certaines caractéristiques ne sont pas normalisées, telles que le nombre de mots (WC) ou le nombre de mots par phrase (WPS). Chaque score est calculé indépendamment pour chaque tweet.

3.3.2 Quantification de la controverse

En se basant sur l'analyse des communautés, la controverse est quantifiée à l'aide de propriétés structurelles et textuelles en examinant deux scores différents.

1. Une description complète des caractéristiques de LIWC peut être trouvée dans [Boy+22].

Score de controverse basé sur la structure du graphe. Un score de controverse basé sur les marches aléatoires (rwc_score) (cf. RWC score dans la section 2.2.2.1) est calculé, en s'appuyant sur les travaux de [Gar+18a] sur le graphe décrit dans la section 3.3.1, en nous concentrant sur le partitionnement des utilisateurs. Ce score a été choisi parce qu'il présente les meilleurs résultats parmi les scores présentés dans [Gar+18a].
Le rwc_score est basé uniquement sur les informations structurelles du graphe, générant de multiples marches aléatoires à partir des nœuds de chaque communauté, et regardant la proportion de marche aléatoire se terminant dans la même communauté que celle d'où elle est partie, en termes de probabilité. Un rwc_score élevé correspond à des communautés séparées, et donc à un sujet plus polarisé et controversé.

Score de controverse basé sur le contenu des tweets. Ce score est basé sur les performances d'un modèle de langage large à pouvoir correctement classer les tweets dans leurs communautés respectives, à partir du contenu textuel des tweets. Nous basons nos travaux sur un modèle de type BERT [Dev+19]. BERT est un modèle d'apprentissage automatique utilisé pour le traitement du langage naturel. Il est basé sur le mécanisme des "Transformers", utilisant plusieurs couches d'attention pour représenter le texte en fonction de son contexte. BERT est pré-entraîné sur un corpus de millions de textes, et peut-être affiné ensuite pour des tâches spécifiques. Nous avons divisé l'ensemble des tweets en deux ensembles d'entraînement et de test, en répartissant équitablement pour chaque ensemble les tweets entre les communautés. L'ensemble de test étant également équilibré, nous utilisons la précision du modèle entrainé sur l'ensemble de test acc_bert comme mesure de performance, et comme score quantifiant la controverse du point de vue du texte. Un score acc_bert élevé correspond à la capacité du modèle à correctement prédire les communautés auxquelles appartiennent les tweets, et donc à un sujet plus controversé.

3.3.3 Explication de la controverse à partir des communautés

Dans cette section, nous présentons l'analyse mise en place afin d'expliquer les communautés présentes sur les sujets jugés controversés. Une analyse descriptive et statistique des caractéristiques conceptuelles des tweets provenant de LIWC selon leurs communautés respectives est d'abord présentée dans la section 3.3.3.1. Une approche autour de l'analyse SHAP des contributions des caractéristiques textuelles des tweets à la classification des communautés utilisateurs est ensuite présentée dans la section 3.3.3.2.

3.3.3.1 Analyse statistique des caractéristiques textuelles générées

Cette analyse est appliquée aux sujets jugés comme controversés. L'analyse statistique a été réalisée à l'aide de Matlab R0021b et de l'outil "Statistics and Machine Learning Toolbox v12.2". La normalité a été testée à l'aide du test d'hypothèse paramétrique de Shapiro-Wilk. Pour tester les différences entre les communautés, l'analyse de la variance (ANOVA) a été utilisée lorsque les hypothèses de l'ANOVA étaient remplies. Dans le cas contraire, le test non paramétrique de Kruskal-Wallis a été utilisé.

La corrélation linéaire entre les variables a été évaluée à l'aide du coefficient de corrélation produit-moment de Pearson. Les corrélations statistiquement significatives sont considérées comme très fortes si $|\rho| \geq 0,8$, comme fortes si $0,5 \leq |\rho| < 0,8$, et faibles dans le cas contraire.

3.3.3.2 Analyse des modèles de classification à l'aide de SHAP

Nous examinons maintenant les sujets présentant des scores rwc_score et acc_bert élevés, considérés comme controversés. Ces scores sont calculés lors de la quantification de la controverse, présentée dans la section 3.3.2.

Cette section participe à l'analyse et à l'explication des caractéristiques qui oriente les modèles lors de la classification des tweets vers une communauté plutôt qu'une autre. Plus le sujet sera jugé comme controversé par nos différents scores, plus l'analyse des communautés sera pertinente car plus le modèle est précis dans ces prédictions, plus les contributions des caractéristiques prennent sens. Nous analysons des tweets provenant de l'ensemble de test et provenant d'utilisateurs des deux communautés, en cherchant à déterminer les caractéristiques textuelles pouvant caractériser chacune des communautés. Afin d'analyser la contribution de chaque caractéristique textuelle aux différents modèles de classification des tweets, nous nous appuyons sur la méthode SHAP.

SHAP [LL17] s'appuie sur la théorie des jeux collaboratifs pour expliquer une prédiction/classification $p(x)$ pour une instance x donnée. La théorie des jeux collaboratifs peut être considérée comme un ensemble de joueurs collaborant afin d'atteindre un objectif commun et diviser équitablement la récompense du jeu. SHAP est une méthode indépendante du modèle. Elle peut être utilisée pour expliquer n'importe quel modèle de prédiction/classification à partir de ses entrées et de ses sorties. L'explication est donnée en termes de contribution marginale de chaque caractéristique de l'instance x à la sortie $p(x)$. Dans notre cas, le modèle de classification des tweets sera considéré comme le jeu, et les caractéristiques du texte des tweets comme les joueurs. Dans ce travail, nous considérons trois modèles de classification de tweet p : BERT, les forêts aléatoires (RF), ainsi qu'un arbre de décision (DT). Pour chaque modèle de classification de tweets, nous nous appuyons sur l'ensemble des caractéristiques textuelles F, comme décrit dans la section 3.3.1. En considérant le même ensemble de test du sujet T créé dans la section 3.3.2, ainsi que le type de caractéristiques étudiées F, nous examinons la contribution marginale de chacune des caractéristiques.

À partir d'une coalition de caractéristiques $(S \in F)$, ne contenant pas la k^{ieme} caractéristique f_k $(f_k \notin S)$, la contribution marginale partielle de la caractéristique f_k pour un tweet donné t et un modèle de classification donné p est calculée comme suit :

$$p\left(t_{S\cup\{f_k\}}\right) - p\left(t_S\right) \tag{3.1}$$

$p(t_S)$ correspond à la prédiction effectuée par le modèle en utilisant uniquement les caractéristiques de la coalition S du tweet t. Toutes les caractéristiques n'appartenant

pas à l'ensemble S sont éliminées. L'équation 3.1 représente l'avantage positif ou négatif que nous obtenons en ajoutant la k^{ieme} caractéristique à la coalition de caractéristiques S. Étant donné un modèle de classification p et ses caractéristiques correspondantes F, la contribution marginale finale d'une caractéristique k^{ieme} f_k de F pour un tweet donné t est représentée par $sv_k(p, F, t)$ et est calculée en considérant toutes les coalitions de caractéristiques possibles S, comme indiqué dans l'équation 3.2. On nomme sv_k la valeur de Shapley de la k^{ieme} caractéristique de F pour un tweet donné.

$$sv_k(p, F, t) = \sum_{S \subseteq F \setminus \{f_k\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \Big[p(t_{S \cup \{f_k\}}) - p(t_S) \Big] \tag{3.2}$$

La valeur de Shapley [LL17] sv est calculée pour chacune des caractéristiques, et pour tous les tweets de l'ensemble de test. Comme indiqué dans l'équation 3.3, le résultat obtenu peut être vu comme une matrice $SV_{p,F}$ dans laquelle $sv_{l,k}$ représente la contribution de la caractéristique f_k pour l'instance de tweet t_l. Chaque ligne horizontale de la matrice SV représente la contribution des différentes caractéristiques au modèle de classification de l'ensemble de tweets correspondant. Chaque ligne verticale représente la contribution d'une caractéristique aux différents modèles de classification. La moyenne des valeurs des lignes verticales peut être considérée comme la contribution d'une caractéristique au modèle pour l'ensemble des tweets d'un sujet. Ainsi, chaque ligne de la matrice fournit l'explication locale d'un tweet, tandis que la matrice entière fournit l'explication globale du modèle sur l'ensemble de tweets.

$$SV_{p,F} = \begin{pmatrix} sv_{11} & sv_{12} & \dots & sv_{1m} \\ \dots & \dots & \dots & \dots \\ sv_{i1} & sv_{i2} & \dots & sv_{im} \\ \dots & \dots & \dots & \dots \\ sv_{n1} & sv_{n2} & \dots & sv_{nm} \end{pmatrix} \tag{3.3}$$

3.4 Préparation des données

Les travaux effectués se basent sur le réseau social Twitter, regroupant une multitude de tweets liés à différents sujets, controversés ou non. Nous appliquons notre analyse sur 30 sujets différents, présenté dans la section 2.4.3. Chaque sujet comprend plusieurs tweets et retweets. Les statistiques sur ces sujets sont regroupés dans la table 2.4. Les tweets sont ensuite préparés et nettoyés au préalable, en remplaçant les URL et les étiquettes des utilisateurs par des jetons spéciaux uniques.

Les caractéristiques textuelles utilisées pour expliquer les communautés étant dépendantes du sujet, nous avons basé la section sur l'explicabilité (section 3.3.3) sur seulement deux sujets à des fins de simplification, l'un étant controversé (PELOSI) et l'autre non controversé (THANKSGIVING). Même si ces deux sujets ont été choisis car présentant des scores élevés de rwc_score et acc_bert, nous avons réalisé des visualisations sur différents sujets pour vérifier la généralisation de nos analyses. Nous présentons les statistiques des deux jeux de données dans la table 3.2.

TABLE 3.2 – *Statistiques descriptives des deux communautés récupérées pour les ensembles de données* PELOSI *et* THANKSGIVING*. La catégorie "Utilisateurs avec tweets*" représente les utilisateurs postant au moins un tweet original.*

-	PELOSI			THANKSGIVING		
	C_0	C_1	Total	C_0	C_1	Total
Tweets	10 430	5 087	15 517	5 531	13 512	19 043
Utilisateurs	48 032	45 230	93 262	55 141	55 138	110 279
Utilisateurs avec tweets*	5 900	3 222	9 122	4 781	10 158	14 939

PELOSI. Il s'agit d'un sujet controversé concernant le discours de Nancy Pelosi au Congrès sur la première destitution de l'ancien président américain Donald Trump, le 19 décembre 2019. Trump est accusé d'abus de pouvoir et d'obstruction au Congrès. Le discours, poussant à la destitution de Trump, est critiqué pour de multiples raisons (complots, illégale, etc.) par les personnes défendant l'ancien président Donald Trump, mais aussi celles opposées aux positions de Nancy Pelosi, notamment celle montrant son opposition à l'avortement. Deux communautés principales sont représentées, la première dite "pro-Pelosi", où les utilisateurs soutiennent Nancy Pelosi. Dans la seconde, dite "anti-Pelosi", les utilisateurs sont soit contre Nancy Pelosi, soit en faveur de Donald Trump. Après avoir effectué le partitionnement des utilisateurs depuis le graphe de retweets présenté dans la section 3.3.1, et vérifié les tweets de manière aléatoire, nous avons noté que la communauté C_0 représente plutôt les personnes opposées à la députée Pelosi, anti-démocrates, tandis que la communauté C_1 comprend des utilisateurs soit pro-Pelosi, soit opposés à Trump, soit favorables à la destitution.

THANKSGIVING. Il s'agit d'un sujet étiqueté comme non controversé rassemblant des tweets faisant référence aux célébrations de Thanksgiving en 2019, fête nationale annuelle aux États-Unis célébrant la récolte et autres bénédictions de l'année écoulée. Du fait de son caractère non controversé, les communautés récoltées après le partitionnement n'ont pas forcément une représentation précise pour le moment.

3.5 Résultats et expérimentations

Nous avons appliqué les trois premières étapes de notre approche sur les 30 sujets présentés dans la section 2.4.3 afin de juger la qualité de la quantification des scores de controverse. Dans la section 3.5.1, nous présentons les conditions de création de nos graphes. Dans la section 3.5.2, nous présentons les résultats et les performances de nos scores de quantification de la controverse. Enfin, la section 3.5.3 porte sur l'analyse des communautés. L'explication de la représentation des communautés restant dépendante et reliée à chaque sujet, cette section ne sera appliquée qu'à deux sujets présentant des scores élevés de rwc_score et de acc_bert mais nous avons relevé des analyses similaires sur d'autres sujets. Les deux sujets, PELOSI et THANKSGIVING, ont été manuellement labellisés respectivement comme controversé et non controversé.

3.5.1 Création du graphe utilisateur

Construction du graphe et partitionnement en communautés. 30 graphes entièrement connectés sont construits à partir des 30 sujets puis chacun est divisé en deux communautés distinctes C_0 et C_1. La proportion d'utilisateurs $(CPROP)$ entre C_0 et C_1 est calculée indépendamment pour chaque sujet selon l'équation 3.4.

$$CPROP = \frac{min(|C_0|, |C_1|)}{max(|C_0|, |C_1|)} \tag{3.4}$$

L'éventail des proportions d'utilisateurs de nos différents graphes est large, et varie de 0,05 à 0,99 avec une moyenne de 0,54. Ces statistiques présentes des différences structurelles évidentes entre les différents sujets considérés.

Traitement du contenu des tweets. Pour chaque sujet, trois ensembles de caractéristiques textuelles différents sont extraits. Les ensembles TF-IDF-TOKENS, BERT-TOKENS et LIWC-FEATURES seront utilisés par nos modèles de classification. Les caractéristiques LIWC sont récupérées à l'aide de l'outil LIWC-app [2] sur chaque tweet indépendamment les uns des autres. Plusieurs sujets étant rédigés dans des langues différentes, chaque tweet est traduit de leur langue originale vers l'anglais, de manière indépendante, en utilisant la bibliothèque python "Deep translator" [3], combinée à l'algorithme de "Google Translator".

Étant donné un sujet t, l'ensemble des tweets est représenté par $X \in G_t$ pour chaque auteur u_i. Chaque tweet est étiqueté avec le label c_i correspondant à celui de son auteur.

3.5.2 Quantification de la controverse

Nous comparons les caractéristiques des 30 sujets indépendamment. De manière à mieux quantifier la qualité des scores, la sensibilité de la précision des modèles est mesurée à l'aide de l'aire sous la courbe ROC (AUC-ROC). Cela permet de mesurer le chevauchement entre les différents scores des sujets controversés et non controversés. Un score AUC-ROC de 1 représente une séparation parfaite entre les sujets controversés (score de controverse proche de 1) et non controversés (score proche de 0), tandis qu'un score de 0,5 indique des communautés indiscernables. L'objectif est de voir si, du point de vue de la communauté, les contenus des tweets, en plus des informations structurelles, peuvent fournir des informations sur la controverse. Il s'agit également de déterminer si les tweets de sujets controversés de chaque communauté sont plus faciles à capturer et à classer par nos modèles.

Score basé sur la structure. En se basant uniquement sur les propriétés structurelles, le rwc_score est calculé pour chacun des 30 sujets. Un score AUC-ROC final élevé de $0,88$ est obtenu sur les rwc_score. Ce résultat montre une bonne séparation entre les sujets. D'après les informations purement structurelles, les sujets controversés montrent

2. https://www.liwc.app/
3. https ://pypi.org/project/deep-translator/

des comportements similaires au niveau des interactions utilisateurs, comparé à ceux des sujets non controversés.

Score basé sur le texte. Pour chaque sujet t, l'ensemble respectif de tweets X est divisé en deux ensembles d'entrainement et de test totalement équilibrés, en utilisant un ratio de 0,8. Pour chaque sujet, et basé uniquement sur les propriétés textuelles, le score acc_bert représente le score de précision sur l'ensemble de test. Concernant le modèle BERT utilisé pour la classification des tweets, nous avons extrait les 12 couches du transformateur et ajouté une couche supplémentaire de 768 neurones avant la couche de classification. Le modèle est entraîné tant que le score de la fonction de perte sur le jeu d'entraînement diminue, avec un maximum de 100 étapes d'apprentissage (ou "epoch"). Un taux d'apprentissage de $2e^{-5}$ est utilisé. Le modèle est optimisé avec l'algorithme d'optimisation "Adam" [4], en utilisant un taux d'apprentissage décroissant pour éviter de perdre trop d'informations des premières couches de transformateurs.

À partir des scores acc_bert obtenus, le score AUC-ROC final est de $0,79$. Un score élevé montre des tweets plus généralisables autour des communautés sur les sujets controversés, représenté par des modèles plus précis. Nous remarquons que certains sujets présentent des déséquilibres significatifs au niveau du nombre d'utilisateurs dans chaque communauté, en particulier pour les sujets non controversés. En ne considérant que les sujets ayant deux communautés fortes avec une proportion d'utilisateurs $CPROP$ supérieure à 0,2 (25 sujets restants), le score AUC-ROC augmente pour les deux scores acc_bert et rwc_score, avec des valeurs montant respectivement à $0,90$ et $0,91$. Enfin, en factorisant les scores rwc_score et acc_bert pour chacun des 30 sujets, nous atteignons une AUC-ROC de $0,91$. Les informations textuelles et structurelles sont donc complémentaires, et leur combinaison améliore l'analyse et la quantification de la controverse. De plus, les scores rwc_score et acc_bert présentent des comportements similaires sur les sujets ambigus.

Les deux méthodes présentent deux scores élevés sur le même sujet non controversé THANKSGIVING, alors que sur le sujet controversé LEADERSDEBATE, ces mêmes méthodes présentent deux valeurs faibles pour les deux scores. Ces expérimentations renforcent notre conclusion selon laquelle le contenu des tweets et les interactions entre utilisateurs contiennent des informations utiles pour quantifier la controverse.

3.5.3 Explication de la controverse à partir des communautés

Dans cette section, la méthodologie présentée dans la section 3.3.3 est appliquée. La contribution des caractéristiques textuelles est examinée, par le biais des prédictions des différents modèles, sur la base des valeurs SHAP calculées pour chaque caractéristique concernée. Pour chaque sujet, nous utilisons la même répartition des utilisateurs sur les ensembles d'entraînement et de test que celle utilisée pour le modèle BERT, décrit dans la section 3.5.2. Enfin, les modèles de forêt aléatoire et d'arbre de décisions sont entrainés, en utilisant à la fois l'ensemble TF-IDF-TOKENS et l'ensemble LIWC-FEATURES (séparément).

4. Extension de la méthode de descente de gradient stochastique.

3.5.3.1 Analyse statistique de la controverse

Une analyse statistique descriptive (corrélation et différences entre les groupes), présentée dans la section 3.3.3.1, a été réalisée sur le sujet controversé PELOSI, afin de mieux comprendre les différences linguistiques entre les deux communautés. Dans cette analyse, nous avons utilisé les caractéristiques conceptuelles LIWC. L'indépendance des observations de l'échantillon a été assurée par deux étapes de prétraitement :

1. Les tweets sont regroupés par utilisateur, puisqu'un utilisateur peut en produire plusieurs, et la valeur moyenne de chaque caractéristique LIWC est calculée, donnant une observation par utilisateur.

2. Tous les utilisateurs participant à plus d'un sujet ont été éliminés du jeu de données. La proportion d'utilisateurs éliminés est inférieure à 7% et est considérée acceptable en comparaison à la quantité totale d'utilisateurs.

Analyse de corrélation. Outre les corrélations positives évidentes existantes entre les variables appartenant à des niveaux hiérarchiques liés (c'est-à-dire ayant des relations parent-enfant), nous avons identifié quelques corrélations statistiquement significatives intéressantes entre les variables. Il convient de noter ici que sur les 101 relations hiérarchiques parent-enfant, seules 17 se sont avérées statistiquement significatives. Une forte corrélation existe entre les variables *Dic - Linguistique* ($\rho = 0,81232, p < 0,001$), indiquant la pertinence des dictionnaires utilisés dans LIWC pour capturer les aspects linguistiques. Une corrélation plus évidente existe entre le comportement "prosocial" ("altruisme", "serviabilité") et la "politesse" : *prosocial - polite* ($\rho = 0,5336, p < 0,001$), bien que ces deux caractéristiques n'appartiennent pas à la même hiérarchie. Une autre corrélation négative intéressante existe entre *Clout* (le langage du leadership, du statut) et *Authentic* (l'honnêteté et l'authenticité perçues) ($\rho = -0,3177, p < 0,001$), ce qui suggère que les utilisateurs qui parlent de leadership et de statut sont moins polis. Enfin, la *tonalité négative* (incluant des notions telles que la méchanceté, le mensonge, la haine) est corrélée à l'*émotion* (incluant des notions telles que la bonté, l'amour, bonheur et espoir), ce qui suggère que ces sentiments opposés coexistent.

Différences entre les groupes. L'analyse des moyennes entre les deux communautés différentes a révélé quelques faits intéressants. Tout d'abord, sur les 117 caractéristiques de l'outil LIWC-22, seules 29 ne présentaient pas de différences statistiquement significatives entre les communautés. Pour le groupe de variables *Summary*, *Analytical thinking*, *Authentic* (honnêteté perçue) et le *Pourcentage de mots de sept lettres ou plus* ne présentaient pas de différences statistiquement significatives entre les deux groupes. En termes de caractéristiques linguistiques, l'utilisation de la 1^{re} personne du singulier (-0,591, $p < 0,001$), de la $3^{ème}$ personne du singulier (0,2338, p=0,014) ainsi que de la $3^{ème}$ personne du pluriel (0,2909, $p < 0,001$) présentent des différences statistiquement significatives entre les communautés. Les mentions de la 1^{re} personne du pluriel ou de la $2^{ème}$ personne ne présentent pas de différences statistiques entre les communautés, les différences se rapportant aux moyennes des communautés $C0$-$C1$. Les variables du groupe *psychological processes* liées à *Cognition* (0,3428, p=0,002), *positive ton* (-0,6473, p<0,001), *negative tone* (0, 7351, p<0,001), *positive emotions* (-0,3333, p<0,001), *anger* (0,1106, p=0,003), *female* (0,439, p<0,001) ou *male* (-0,3433, p<0, 001) ont des moyennes différentes statistiquement significatives entre les deux communautés. Les variables

TABLE 3.3 – *Présentation des variables LIWC ayant des différences statistiquement significatives, lors de l'analyse des moyennes sur le sujet controversé* PELOSI.

Catégories	Différences	Pas de Différences
SUMMARY DIMENSION	*Summary, Analytical thinking Authentic, BigWords*	*Tone, Clout, WC*
LINGUISTIC	*I, He, Them*	*We, You*
PUNCTUATION MARKS	*Question, Exclamation, Apostrophes*	*Period, Coma*
PSYCHOLOGICAL PROCESSES	*Cognition, positive ton, negative tone, positive emotions, anger, female, male*	*Insights, Differentiation, Emotion, Anxiety, Sadness, Prosocial behavior, Interpersonal Conflict, Moralization*
EXPANDED DICTIONARY	*Politics, Ethnicity, Lifestyle, Religion, Physical status, Sexuality, Death, Past, Present, Future*	*Technology, Home, Acquire, Fatigue, Curiosity, Allure, Attention, Space, Feeling, Non-fluencies*

relatives à *Insights*, *Differentiation*, *Emotion*, *Anxiety*, *Sadness*, *Prosocial behavior*, *Interpersonal Conflict*, ou *Moralization* n'ont pas de différences statistiquement significatives entre les communautés. Dans la catégorie *Expanded Dictionary*, les caractéristiques relatives à *Politics* (0,0179, p=0,002), *Ethnicity* (0,2971, p<0, 001), *Lifestyle* (0,2361, p=0,001), *Religion* (0,53895, p<0,001), *Physical status* (par exemple, médicaments, alimentation, santé, maladie, etc.) (0,62998, p<0, 001), *Sexual* mentions (0,1126, p<0,001), ou *Death* (0,073, p<0,001) ainsi que les caractéristiques reflétant l'intérêt de l'utilisateur pour le passé (-0,5325, p<0,001), le présent (0,5382, p<0,001) ou l'avenir (0,2594, p<0,001) présentent des différences statistiquement significatives entre les deux communautés. Au contraire, les caractéristiques qui ne présentent pas de différences moyennes statistiquement significatives entre les communautés comprennent les variables liées à *Technology*, *Home*, *Acquire*, *Fatigue*, *Curiosity*, *Allure*, *Attention*, *Space*, *Feeling*, et *Non-fluencies*, indiquant que ces caractéristiques ne diffèrent pas entre les deux populations. Enfin, les signes de ponctuation tels que l'utilisation de *Question* (0,30206, p<0,001) ou de *Exclamation* (0,4118, p<0,001) ainsi que l'utilisation de *Apostrophes* (-0,543, p<0,001) présentent des différences statistiquement significatives entre les deux communautés. Le tableau 3.3 regroupe les variables en deux catégories selon la significativité de leur différence au niveau des moyennes.

Nous étudions maintenant l'étape d'explicabilité des communautés de notre approche, présentée dans la section 3.3.3, sur le même jeu de données PELOSI, ainsi que sur un jeu de données non controversé présentant des scores de quantification élevés, THANKSGIVING, présenté dans la section 3.4.

3.5.3.2 Analyse communautaire d'un sujet controversé

PELOSI, étiqueté controversé, présente un score $rwc\_score = 0,70$, un score $acc\_bert = 0,79$ et un score de $0,59$ lorsque l'on combine ces deux scores. Comme le montre la

TABLE 3.4 – *Mesure de précision pour différentes combinaisons de modèles et de caractéristiques appliquées sur $test_{pelosi}$, jeu de données de test sur le sujet PELOSI, pour la tâche de classification des tweets dans leur communauté correspondante.*

Type de Modèle	Type de caractéristique	Identifiant	**Précision**
DECISION-TREE	TF-IDF	DT_{tfi}	0,65
	LIWC	DT_{liwc}	0,62
	TF-IDF + LIWC	$DT_{tfi+liwc}$	0,66
RANDOM-FOREST	TF-IDF	RF_{tfi}	0,69
	LIWC	RF_{liwc}	0,68
	TF-IDF + LIWC	$RF_{tfi+liwc}$	0,71
BERT	TEXT	$BERT_{text}$	**0,79**

figure 3.2, nous remarquons deux communautés distinctes, où les utilisateurs sont fortement liés l'un à l'autre tout en étant moins liés à l'autre communauté, ce qui explique le score élevé de rwc_score.

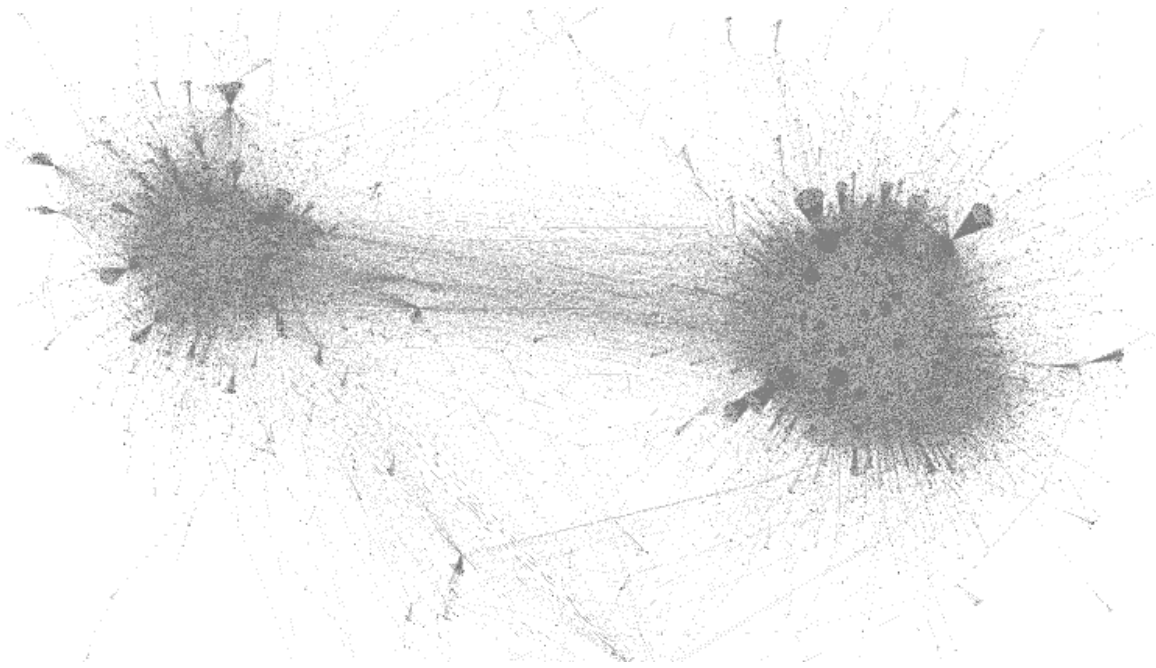

FIGURE 3.2 – *Graphe utilisateur, basé sur les retweets, d'un sujet controversé autour du discours de Nancy Pelosi, le 5 décembre 2019. Le graphe est représenté à l'aide de l'algorithme de visualisation spatiale Force Atlas 2* [Jac+14].

Le tableau 3.4 regroupe les résultats des expériences appliquées à ce sujet. Notre modèle basé sur BERT distingue les tweets provenant d'utilisateurs appartenant à différentes communautés avec une précision de 0,79 et dépasse les performances des modèles DT et RF utilisant de simples caractéristiques statistiques de mots (TF-IDF). Ces résultats montrent clairement que le texte contient des informations importantes sur l'analyse des communautés dans le cas de la controverse.

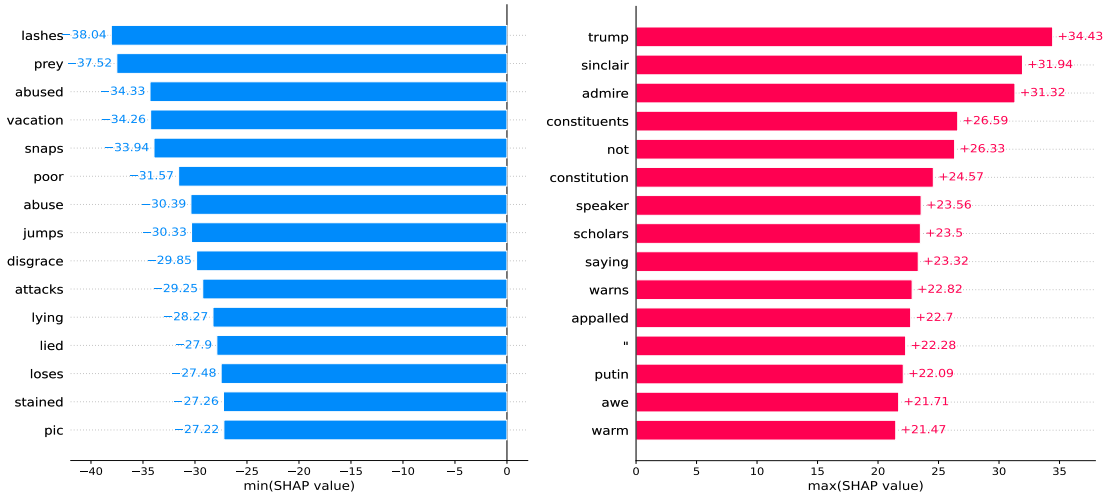

FIGURE 3.3 – *Les valeurs SHAP sont calculées à partir des tweets du jeu de données $test_{pelosi}$, en utilisant le modèle BERT$_{text}$. Les valeurs correspondent à l'impact des tokens (généré sur le tweet) sur la prédiction d'une communauté. La figure montre les 10 premiers tokens ayant le plus d'impact sur la prédiction des communautés C_0 (à gauche) et C_1 (à droite).*

L'analyse des tokens (mots) ayant une contribution importante sur la classification du modèle BERT$_{text}$ renforce notre conclusion, selon laquelle le modèle BERT$_{text}$ a bien capturé les caractéristiques liées aux communautés. La figure 3.3 montre les tokens ayant le plus d'impact dans la prédiction des communautés pour les tweets appartenant à l'ensemble $test_{pelosi}$. Comme attendu, les mots à connotation négative et à tendance péjorative (*abuse*, *disgrace*, *lying*, *loses*, *stained*) poussent fortement le classifieur à prédire que le tweet appartient à un utilisateur de la communauté C_0 attaquant la sénatrice Pelosi. D'autres mots lié à cette communauté mettent également l'accent sur les conspirations (*lashes*, *snaps*, *attacks*), probablement contre Donald Trump. Au contraire, les mots représentant des adjectifs qualificatifs positifs (*admire*, *warm*, *awesome*, *speaker*) ont tendance à influencer fortement le modèle vers la communauté C_1, celle contenant des utilisateurs défendant Nancy Pelosi. Il convient de noter que les mots spécifiques au sujet peuvent également être représentatifs des arguments potentiels d'une communauté. Parmi ces tokens, on peut citer "constitution" (utilisation des lois pour demander la destitution de Donald Trump) ou encore "Sinclair", média ayant attiré la colère de certaines personnes pour avoir soulevé une question ayant pour objectif de mettre Nancy Pelosi dans l'embarras. Enfin, on observe que les utilisateurs de la communauté C_1, du moins par rapport à C_0, ont plus tendance à tweeter ou retweeter en utilisant le guillemet ", souvent utilisé pour citer les pensées d'autres auteurs. Les mots les plus impactants du jeu d'entrainement, regroupés dans la figure 3.4, sont également analysés, afin de comprendre la manière dont le modèle apprend à prédire correctement les communautés. Nous avons constaté que les champs lexicaux des mots autour des communautés sont très proches de l'analyse faite précédemment sur le jeu de

test. Cela montre qu'une manière différente de communiquer existe entre ces deux communautés, via des champs lexicaux distincts.

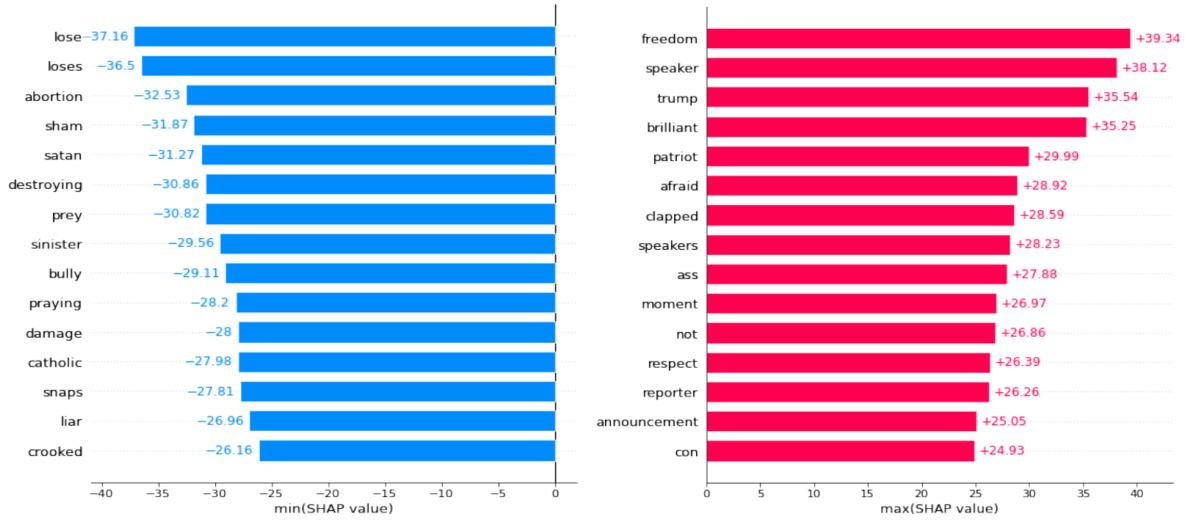

FIGURE 3.4 – *Les valeurs SHAP sont calculées à partir des tweets du jeu de données d'entrainement $train_{pelosi}$, en utilisant le modèle BERT$_{text}$. Les valeurs correspondent à l'impact des tokens (généré sur le tweet) sur la prédiction d'une communauté. La figure montre les 10 premiers tokens ayant le plus d'impact sur la prédiction des communautés C_0 (à gauche) et C_1 (à droite).*

En ce qui concerne les caractéristiques LIWC relatives aux états psychologiques, nous obtenons une précision de 0,68 sur le modèle RF$_{liwc}$, atteignant même 0,71 lorsqu'elles sont combinées avec les caractéristiques TF-IDF relatives aux mots. Sur la base de ces résultats, nous pouvons supposer que dans ce cas controversé, les caractéristiques LIWC peuvent aider à caractériser une tendance dans une communauté par rapport à une autre.

Dans l'objectif de trouver des tendances aux niveaux des états psychologiques dans les communautés [5], la figure 3.5 montre les principales caractéristiques LIWC ayant un impact sur la prédiction du modèle RF$_{liwc}$ appliqué sur l'ensemble de données $test_{pelosi}$. Ces caractéristiques appartiennent aux catégories *Tone*, *function*, *Period*, *Exclam*, *OtherP*, *Cognition*, *Affect*, *Social*, *Lifestyle*, *Physical* et *Time orientation*, présentées dans le tableau 3.1. Nous remarquons que la ponctuation joue un rôle important (*Exclam*, *OtherP*, *period*). Le token " *!*" par exemple montre un impact élevé sur la prédiction de la communauté C_0, ce qui est cohérent, puisque les utilisateurs attaquant Pelosi utilisent généralement des sentiments forts ou de l'emphase dans leurs tweets. La figure 3.5 indique également que des fonctions comme les pronoms (*I*, *They*) ou les chiffres ont un impact sur les prédictions du modèle. Le pronom *They* a tendance à avoir un impact positif sur les prédictions de C_0, en comparaison à la première personne du singulier (*I*). Nous pouvons également accorder une attention particulière au ton et aux émotions ressenties par chaque communauté. Nous remarquons que le ton est une caractéristique très impactante et discriminante du modèle. Comme

5. Une caractéristique dite *parfaite* représenterait deux groupes de couleurs bien séparés, loin de la frontière de décision (abscisse à 0).

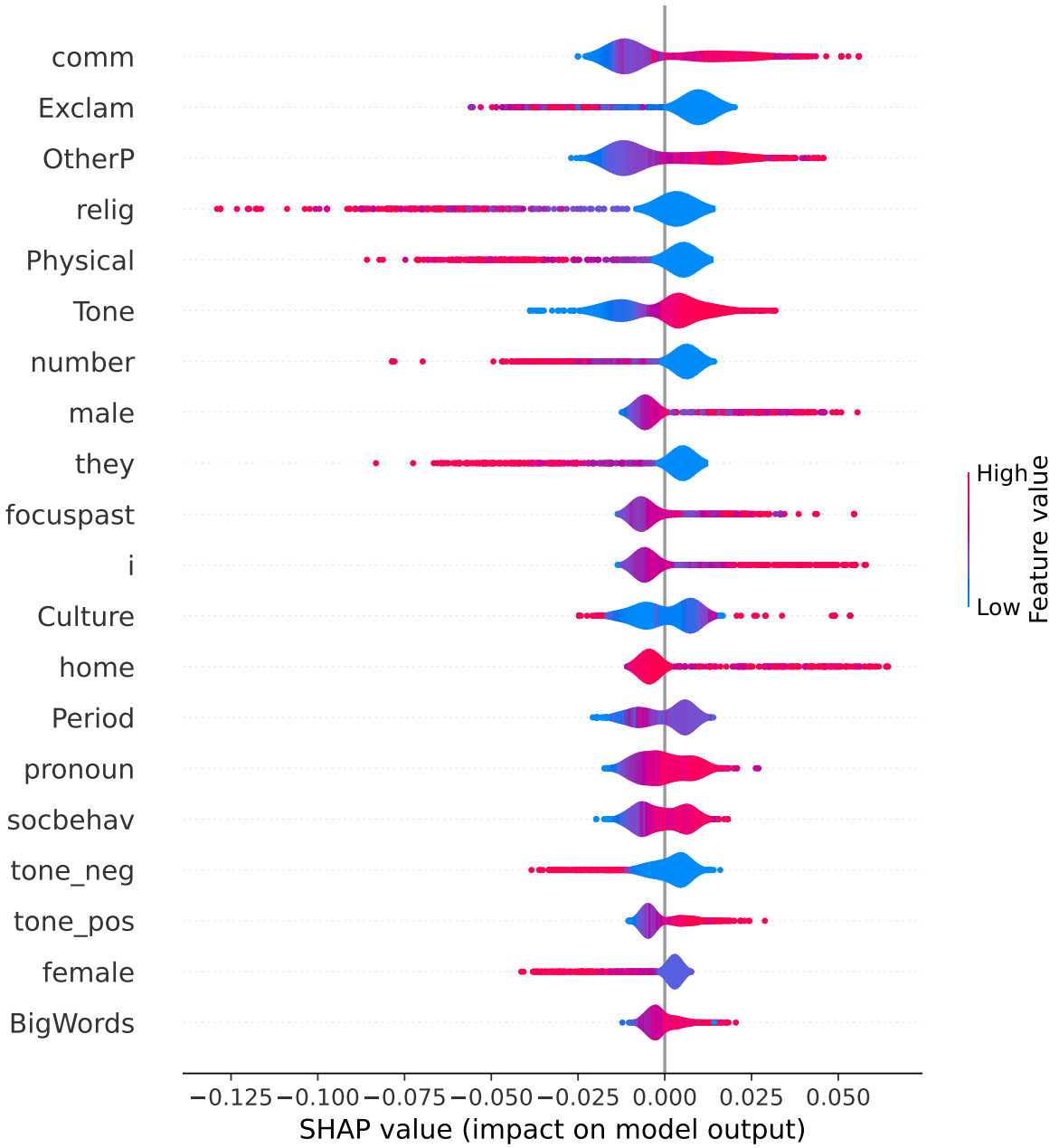

FIGURE 3.5 – *Projection des 20 caractéristiques LIWC ayant le plus d'impact sur les prédictions du modèle* $textscr f_{liwc}$ *sur l'ensemble* $test_{pelosi}$*. L'échelle de couleurs, du rouge (faible valeur de la caractéristique) au bleu (valeur élevée) est représentée pour chaque échantillon. Plus la valeur absolue de SHAP est grande, plus la caractéristique pousse le modèle à prédire le tweet vers* C_1 *(ou vers* C_0*, selon le signe de la valeur).*

indiqué sur les courbes plutôt inversées de *tone_pos* et *tone_neg*. Les utilisateurs de la communauté C_1, soutenant Nancy Pelosi, sont plus susceptibles d'utiliser un ton positif que la communauté C_0, employant généralement un ton plus dramatique ou polémique. Cela correspond à nos conclusions concernant l'analyse de BERT$_{text}$ faite précédemment. Enfin, nous remarquons que les variables *home*, *period* et *BigWords*, n'ont statistiquement aucune différence entre les communautés (selon l'analyse statistique faite précédemment), mais sont toujours identifiées comme des contributeurs majeurs du classifieur (figure 3.5), ce qui montre que SHAP identifie des comportements à partir de caractéristiques différentes de celles identifiées par l'analyse statistique.

Pour conclure, concernant la proportion d'utilisateurs et de tweets de chaque communauté du sujet PELOSI, le tableau 3.2 montre que dans ce sujet controversé, la communauté dite "attaquante" (celle en désaccord avec le propos du sujet) (C_0) est plus importante que la communauté dite "défendante" (C_1). Ceci n'étant qu'une interprétation partielle et simplifiée, d'autres analyses pourraient être développées à partir de cette analyse d'impact des caractéristiques autour de ce sujet controversé, aidant ainsi à la compréhension globale des diverses communautés.

3.5.3.3 Analyse communautaire d'un sujet non controversé

Le sujet non controversé THANKSGIVING présente des scores $rwc\_score = 0,78$ et $acc\_bert = 0,74$, ainsi qu'un score de $0,58$ lorsque l'on factorise ces deux scores. Ce sujet montre deux communautés fortes (la proportion $CPROP$ est supérieure à 0,2) tout en étant étiqueté comme non controversé. Ce sujet a été choisi pour être étudié, afin de comprendre ce qui fausse la quantification des deux scores de controverse, notamment le modèle basé sur BERT afin de prédire les communautés correctes des tweets.

En visualisant le graphe de retweets, à l'aide du même algorithme de présentation forcée que celui utilisé pour PELOSI dans la figure 3.2, nous remarquons que la communauté C_1 a des utilisateurs extrêmement proches les uns des autres, tandis que la communauté C_0 a des utilisateurs plus éloignés. Cela explique un score excessivement élevé de rwc_score. Cependant, les deux communautés ne semblent pas très éloignées, comparées au topic PELOSI. Ensuite, les expériences réalisées à l'aide du modèle BERT$_{text}$ sur l'ensemble de test donnent une précision de 0,74 (correspondant au score acc_bert). En entrainant une forêt aléatoire avec les caractéristiques LIWC (RF$_{tfi+liwc}$) sur le même jeu de test, on obtient une précision de 0,70. Sur la base de la même analyse que celle présentée dans la section 3.3.3, nous analysons la contribution des caractéristiques ayant un impact sur la classification des tweets dans les communautés. La figure 3.6 montre les caractéristiques ayant le plus d'impact sur la prédiction de chaque communauté à l'aide du modèle basé sur BERT. Nous remarquons que si la communauté C_0 contient des mots (tokens) n'appartenant pas nécessairement à une catégorie commune, la communauté C_1 contient sept mots liés à la politique (par exemple, "président", "politique", "Trump"). Les utilisateurs de la communauté C_1 semblent parler davantage de politique (tout en étant fortement liés les uns aux autres), suggérant que le sujet pourrait être lié à un sous-sujet controversé. La communauté C_0, au contraire, utilise un vocabulaire plus souple, sans rassembler les utilisateurs sur un domaine particulier. Cela peut expliquer la grande capacité du modèle à effectuer

des tâches de classification des tweets dans les bonnes communautés, par rapport à d'autres sujets non controversés. On remarque aussi que la caractéristique *politic* fait partie des 20 caractéristiques les plus impactantes (selon l'analyse SHAP), par notre modèle $\text{RF}_{tfi+liwc}$, correspondant à ce que nous avons retenu de l'analyse précédente.

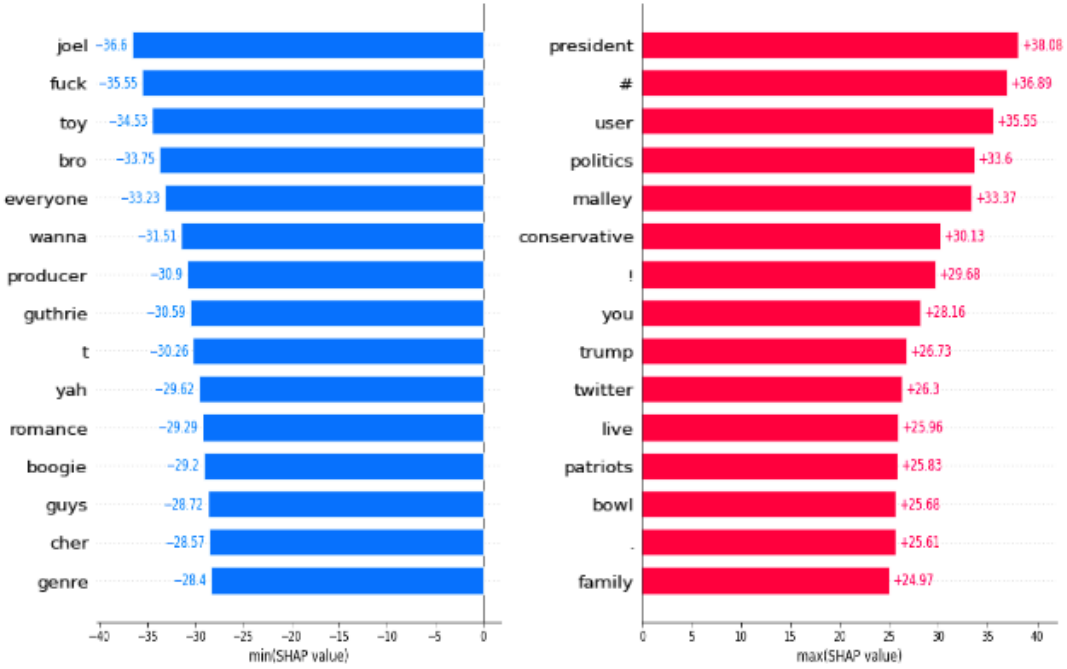

FIGURE 3.6 – *Les valeurs SHAP sont calculées à partir des tweets du jeu de données $test_{thanksgiving}$, en utilisant le modèle BERT_{text}. Les valeurs correspondent à l'impact des tokens (générés sur le tweet) sur la prédiction d'une communauté. La figure montre les 10 premiers tokens ayant le plus d'impact sur la prédiction des communautés C_0 (à gauche) et C_1 (à droite).*

3.6 Conclusion

Nous avons présenté dans ce chapitre une chaîne de traitements analysant la controverse sur Twitter afin de quantifier les sujets controversés et d'expliquer la controverse d'un point de vue des communautés d'utilisateurs. Nous nous sommes appuyés sur l'utilisation de différents ensembles de caractéristiques textuelles et conceptuelles, ainsi que sur la méthode SHAP afin d'identifier les caractéristiques contribuant à la classification des tweets dans leurs communautés via différents modèles. Les expériences menées montrent que l'explication des communautés fonctionne bien sur les sujets identifiés comme controversés, et donc ayant des scores rwc_score et acc_bert élevés, même si certains sujets non controversés peuvent également avoir des communautés structurées facilement identifiables. Cela confirme qu'outre le fait que la controverse soit une notion subjective, elle devrait être considérée de manière floue et non binaire. Sa quantification permet aux utilisateurs de mieux comprendre dans quelle mesure un sujet est controversé. De plus, cette analyse montre que le texte présente également des caractéristiques intéressantes et complémentaires aux interactions des utilisateurs pour analyser les sujets controversés. Cette étude est basée sur 30 sujets, les expérimentations ne sont pas suffisantes pour généraliser ses résultats. Néanmoins, nous avons proposé

une chaîne de traitements généralisable, afin d'analyser la controverse du point de vue des communautés et avons montré la corrélation de certaines tendances sur des sujets controversés. Notre interprétation est aussi basée sur des labels utilisateurs dits faibles, même si la méthode de partitionnement a récemment montré de bons résultats [Gar+18a].

L'analyse d'un sujet spécifique devenant largement controversé par la compréhension des communautés peut permettre d'améliorer de nombreuses tâches de recherche, telles que l'étude et la quantification des sujets controversés d'un point de vue de leur temporalité. Une autre perspective intéressante pourrait être de généraliser cette approche sur différents médias sociaux. L'inclusion des sous-communautés dans l'analyse reste également une tâche difficile. Ces travaux peuvent contribuer à de futures recherches sur l'amélioration des mesures de détection et quantification de la controverse sur Twitter, en intégrant des données textuelles significatives à des informations structurelles sur des modèles plus complexe, tels que les réseaux de neurones graphiques (GNN). Dans cette optique, nous présenterons dans le chapitre 4, une nouvelle approche, afin de détecter de manière binaire si un fil de discussion Reddit est controversé ou non, en utilisant les GNN afin de faire ces prédictions.

CHAPITRE 4

DÉTECTION DES POSTS CONTROVERSÉS : UNE APPROCHE BASÉE SUR LES RÉSEAUX DE NEURONES, APPLIQUÉE AUX GRAPHES ET AUX TEXTES

Sommaire

4.1 Introduction

Dans ce chapitre, nous présentons les travaux effectués sur la détection de discussions controversées (posts) sur le réseau social Reddit à partir des interactions entre utilisateurs. L'approche est basée sur la classification des posts. Elle utilise les informations textuelles des commentaires (post inclus) et structurelles du graphe utilisateur afin de classer ces posts dans deux catégories, controversés ou non, en se basant sur les GNN comme modèles d'apprentissage. Nous appliquons ensuite notre approche sur un autre réseau social, Twitter, présentant dans ces graphes controversés des propriétés structurelles différentes, afin de juger les capacités de généralisation de notre approche. Ces travaux ont mené à une publication dans la conférence WISE, paru en 2021 [Ben+21], ainsi qu'un article court à la conférence PFIA paru en 2023 [Ben+23a]. Ces travaux ont également abouti à un article étendu dans le journal WWW paru en 2023 [Ben+23b].

Le reste du chapitre est organisé comme suit. La section 4.2 introduit le contexte des travaux et leurs motivations, ainsi que notre approche et les différentes contributions réalisées. La section 4.3 présente une vue d'ensemble de notre approche pour détecter automatiquement la controverse sur les médias sociaux. Ses différentes étapes sont décrites et formalisées. La section 4.4 présente les expériences et les résultats de notre approche sur le réseau social Reddit, ainsi que les discussions autour les résultats obtenus. La section 4.5 présente les expériences menées sur le réseau social Twitter, avec des communautés polarisées lors des sujets controversés. Enfin, la section 4.6 recense les travaux effectués et conclut le chapitre.

4.2 Contexte

Les opinions exprimées dans les médias sociaux suscitent souvent la controverse. Un contenu controversé est un contenu qui suscite des opinions et des interrogations différentes, impliquant des discussions entre utilisateurs de différentes communautés. Ces interactions peuvent prendre la forme de débats ou de discussions par exemple. L'identification de ces sujets controversés de manière automatique reste une tâche difficile. La plupart des approches existantes s'appuient sur la structure du graphe des discussions et/ou du contenu des messages (post et commentaires confondus), mais n'exploitent pas en profondeur les avancées récentes en matière de réseaux neuronaux graphiques (GNN) pour prédire si une discussion est controversée ou non. Les travaux présentés dans ce chapitre visent à combiner les interactions des utilisateurs présentes dans le graphe représentant la discussion et les caractéristiques du texte de la discussion pour détecter la controverse.

Dans ce chapitre, nous nous concentrons sur la détection des posts controversés sur le réseau social Reddit, même si tout autre réseau social (Twitter, Facebook, etc.) peut également être utilisé, après quelques adaptations lors de l'étape de construction du graphe. L'originalité de notre approche réside tout d'abord dans l'utilisation de méthodes GNN très récentes afin de représenter les nœuds (utilisateurs) du graphe dans un espace euclidien à faible dimension, en tenant compte des informations structurelles. Les représentations textuelles initiales des utilisateurs sont apprises à partir

de leurs messages respectifs sur le post concerné, puis utilisées comme entrée de notre modèle basée sur les GNN. À notre connaissance, seuls Zhong et al. [Zho+20] utilise les GNN, en construisant leur méthode de détection des posts controversés à partir de la structure de l'arbre des commentaires [Zho+20]. Nos travaux, au contraire, exploitent le graphe des interactions entre utilisateurs construit à partir de la structure de l'arbre des commentaires, tout en comparant différentes approches autour des GNN pour combiner à la fois les informations structurelles et textuelles.

Contributions. Dans ce chapitre, nous nous intéressons à la détection de la controverse sur les réseaux sociaux. Afin de détecter les posts controversés, nous proposons une approche basée sur les GNN, dont les principales contributions sont :

- **Détection de la controverse**. Nous proposons une approche basée sur les GNN pour la détection de posts controversés, en se basant sur une tâche de classification de graphes. Nous proposons deux stratégies pour représenter l'ensemble de la structure du graphe. La première stratégie vise à exploiter la structure hiérarchique pouvant exister dans le graphe utilisateur. Les informations du graphe sont agrégées sur les arêtes de manière itérative et hiérarchique. Dans notre travail, nous nous appuyons sur l'approche DIFFPOOL, encodant l'ensemble du graphe en empilant plusieurs couches de regroupement [Yin+18]. La deuxième stratégie est basée sur le mécanisme d'attention. Elle vise à permettre à chaque nœud utilisateur de juger quel nœud voisin est plus ou moins important que les autres, pendant le processus d'encapsulation des nœuds, en fonction de la structure et des caractéristiques du graphe.
- **Étude expérimentale**. Nous menons des expériences sur des jeux de données réelles provenant du réseau social Reddit, afin d'évaluer l'approche proposée basée sur les GNN, en utilisant d'abord uniquement les informations structurelles. Nous montrons que notre approche obtient de bonnes performances par rapport aux travaux proposés dans la littérature.
- **Caractéristiques textuelles**. Nous montrons qu'incorporer la représentation textuelle initiale des utilisateurs peut améliorer les performances de notre méthode.
- **Généralisation.** Nous appliquons enfin notre approche sur un réseau social présentant des propriétés différentes (Twitter) afin d'étudier la capacité de généralisation de notre approche.

4.3 Méthodologie

Cette section décrit notre approche concernant la détection de la controverse au niveau des posts. L'idée principale est d'exploiter à la fois le contenu textuel et les interactions entre utilisateurs en représentant la discussion Reddit sous la forme d'un graphe d'utilisateurs et en explorant des techniques avancées de représentation du graphe. Notre méthode est basée sur les réseaux de neurones graphiques GNN pour améliorer la représentation des caractéristiques structurelles du graphe. La figure 4.1 présente une vue d'ensemble de notre approche. Nous divisons notre pipeline en

quatre étapes séquentielles décrites dans la figure 4.1 : (1) Construction du graphe, (2) Extraction des caractéristiques de l'utilisateur, (3) Représentation du graphe et (4) Classification du graphe.

L'étape de construction du graphe représente les données extraites d'un post sous la forme d'un graphe d'utilisateur. Nous représentons l'arbre des commentaires initiaux sous la forme d'un graphe Reddit. Les nœuds représentent les utilisateurs et les arêtes correspondent aux interactions qui existent entre les utilisateurs. Chaque nœud peut être représenté par différentes caractéristiques de l'utilisateur, telles que l'identifiant, l'âge, la localisation, les textes, etc. Dans notre cas, seul le contenu des commentaires/posts de chaque utilisateur est utilisé.

L'étape d'extraction des caractéristiques de l'utilisateur permet d'enrichir les nœuds du graphe avec des informations sur le contenu proposé par l'utilisateur, en ajoutant une représentation vectorielle des caractéristiques textuelles de l'utilisateur. Ces caractéristiques sont calculées à l'aide de méthodes de traitement automatique du langage (TAL) basées sur les réseaux de neurones. Cela permet de mieux interpréter le contenu des textes envoyés par les utilisateurs que les modèles classiques de TAL.

L'étape de représentation du graphe calcule la nouvelle représentation de l'ensemble du graphe dans un espace vectoriel à faible dimension. Différentes techniques avancées d'apprentissage de représentations des graphes basées sur les GNN sont utilisées, à savoir DIFFPOOL [Yin+18], GCN [KW17] et GAT-GC [ZX20]. Enfin, l'étape de classification du graphe prédit de manière binaire si cette nouvelle représentation du graphe associée au post est controversée ou non.

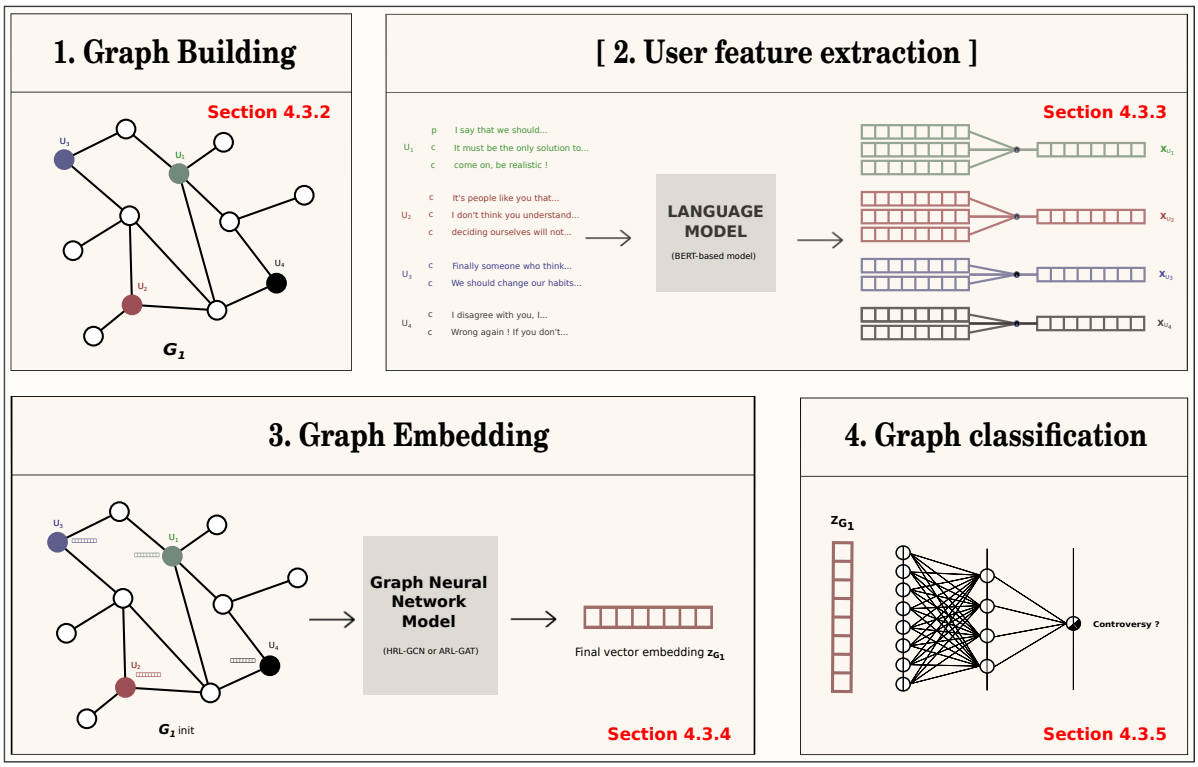

Figure 4.1 – *Vue d'ensemble des quatre étapes de notre approche de détection de la controverse.*

4.3.1 Détection de la controverse : définition du problème

Un sujet controversé désigne tout contenu qui suscite des réactions à la fois positives et négatives. Étant donné que les données relatives à une discussion sont représentées sous forme de graphe (graphe de textes, graphe d'utilisateurs), la détection des posts controversés vise à classer tout nouveau graphe sur la base de graphes déjà étiquetés comme controversés ou non controversés. Plus formellement, nous définissons le problème de la détection de controverse comme suit :

Definition 1. *Soit G un ensemble de graphes, et $G^L = \{(G_1, l_1), ..., (G_i, l_i), ...(G_n, l_n)\}$ un ensemble de graphes étiquetés : $G_i \in G$ est une représentation graphique d'une discussion donnée, et l_i son label correspondant (1 si controversé, 0 sinon). Le problème de détection de la controverse est défini comme un problème de classification, et notre objectif est d'apprendre une fonction de correspondance $f : G \to L = \{0, 1\}$ qui attribue un label (0 ou 1) à chaque graphe non étiqueté sur la base d'un entraînement effectué sur des graphes étiquetés G^L.*

La table 1, présente dans la section 6.2.3, retranscrit la nouvelle formalisation utilisée dans ce chapitre, en faisant le lien avec celle introduite dans notre état de l'art (section 2.1).

4.3.2 Construction du graphe

Les méthodes existantes de détection des posts controversés [HL19 ; Zho+20] sur Reddit utilisent la représentation classique de l'arbre des commentaires et des messages car elles se concentrent principalement sur la structure de la discussion. Cependant, de nombreux travaux de recherche ont établi que l'interaction entre utilisateurs peut être utile pour extraire différentes caractéristiques dans les médias sociaux pouvant améliorer la détection de la controverse. Dans ce travail, nous adoptons une représentation graphique d'une discussion mettant en évidence ces interactions entre utilisateurs.

Étant donné une discussion sur un post p extrait d'un subreddit s, nous construisons un graphe non dirigé où un nœud u_i représente un utilisateur impliqué dans la discussion. Une arête (u_i, u_j) est créée lorsqu'un utilisateur u_j répond au post p ou à un commentaire posté par un utilisateur u_i. Il convient de mentionner que notre représentation graphique n'autorise pas la répétition des arêtes. Lorsque les utilisateurs u_i et u_j se répondent l'un à l'autre, une seule arête non orientée est représentée. En effet, certaines recherches que nous avons menées (visualisation de la structure arborescente, comparaison textuelle des post-commentaires et des commentaires) ont montré que la différenciation des différentes arêtes entre deux nœuds n'ajoute pas nécessairement plus d'informations.

Plus formellement, un post p est représenté par un graphe $G = (\mathcal{U}, \mathcal{E}, X)$ dans lequel $\mathcal{U} = \{u_1, u_2, ..., u_n\}$ désigne les nœuds utilisateurs, $\mathcal{E} = \{(u_i, u_j)\}_{1 \leq i,j \leq n}$ désigne les arêtes du graphe, et $X \in \mathbb{R}^{n \times e}$, e étant la dimension de la caractéristique, désigne la matrice des caractéristiques des nœuds d'utilisateurs.

Chaque nœud correspond à un utilisateur unique. Il existe une arête entre deux nœuds s'il existe des liens d'interaction entre les utilisateurs correspondants. Le calcul de la matrice X est décrit dans la section 4.3.3.

4.3.3 Extraction des caractéristiques utilisateurs

Afin d'apporter des informations complémentaires à la représentation du graphe, les caractéristiques d'un utilisateur sont extraites des posts (commentaire et/ou post) en utilisant des techniques avancées de TAL. Récemment, différents modèles de langage TAL pré-entraînés sur un grand corpus ont été proposés pour améliorer la représentation dynamique du texte, tels que les modèles BERT [Dev+19]. Ces modèles, basés sur le mécanisme d'attention, représentent chaque mot sous la forme d'un vecteur selon le contexte de contenu dans le reste du texte envoyé. L'extraction des caractéristiques est effectuée pour chaque utilisateur comme suit.

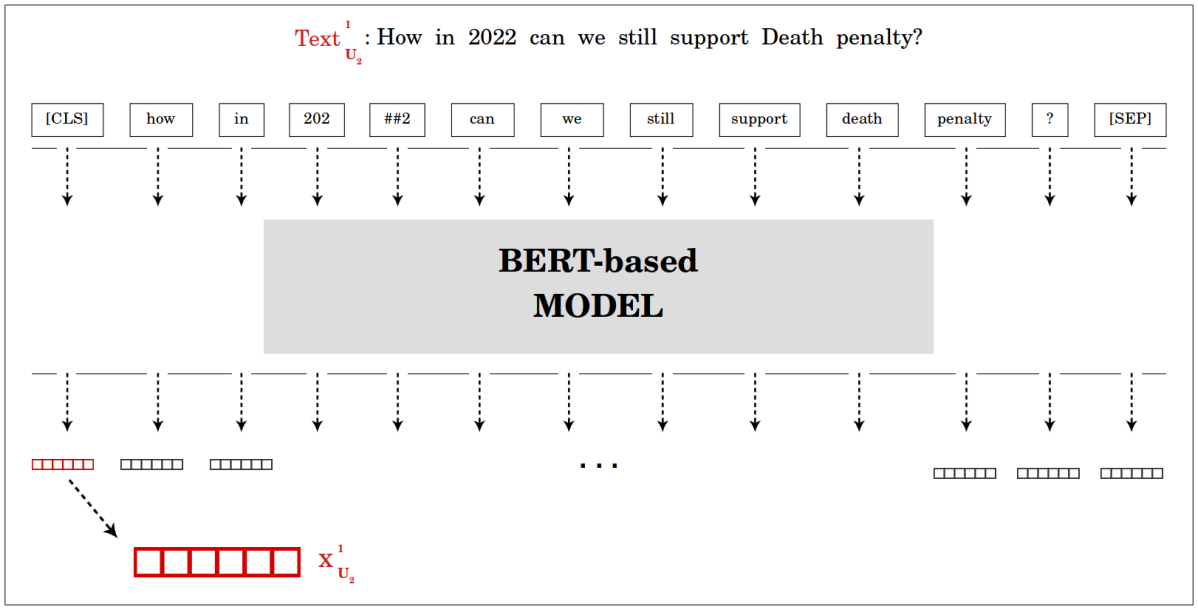

FIGURE 4.2 – *Exemple de représentation d'un commentaire c d'un utilisateur u_2 dans un vecteur de dimension e.*

Chaque message (post ou commentaire) qu'un utilisateur u_i publie est d'abord nettoyé (les tags Reddit et liens URL sont supprimés). Il est ensuite représenté dans un vecteur e-dimensionnel à l'aide d'un modèle de langage BERT pré-entrainé.

Ensuite, comme le montre la figure 4.2, le texte est divisé en différents jetons, à l'aide d'un "tokenizer" spécifique [1] avec une liste de vocabulaire déjà établie, et des jetons spéciaux [CLS] et [SEP]. Le jeton [CLS] représente le jeton de classification, celui utilisé lors les tâches de classification faites à la base de ce modèle. Ensuite, tous les jetons passent par le modèle linguistique basé sur BERT avec plusieurs couches d'auto-attention [Dev+19] et renvoient un vecteur intégré e-dimensionnel pour chacun de ces jetons. Sur la base des travaux effectués dans la littérature, nous utilisons le vecteur de représentation du jeton [CLS] comme représentation vectorielle du message.

Les vecteurs de représentation obtenus à partir des différents messages postés par un utilisateur u_i sont agrégés pour former les caractéristiques finales de l'utilisateur x_{u_i}, comme indiqué dans l'équation 4.1.

1. Nous avons utilisé le tokenizer 'bert-base-uncased' stocké par https://huggingface.co/

$$x_{u_i} = \textsc{Aggregation}\left([x_{u_i}^0, x_{u_i}^1, ..., x_{u_i}^m]\right) \tag{4.1}$$

L'agrégation de ces vecteurs est effectuée en prenant la valeur maximale de chaque dimension (fonction d'agrégation MAX), mais toute autre fonction d'agrégation peut être utilisée.

4.3.4 Représentation du graphe

L'étape de représentation du graphe vise à encoder l'ensemble du graphe utilisateur dans un vecteur à faible dimension. Ce dernier alimente l'étape de classification des graphes pour prédire si le post représenté est controversé ou non.

Récemment, différentes approches basées sur les GNN ont été proposées pour adapter les architectures d'apprentissage profond aux données structurées à des données non structurées telles que les graphes [KW17 ; Vel+18]. L'idée principale est de considérer chaque nœud du graphe comme un nœud de calcul, et d'apprendre les primitives classiques des réseaux de neurones calculant la nouvelle représentation des nœuds.

Cette étape repose sur des architectures GNN dont l'objectif est d'exploiter à la fois les caractéristiques des nœuds de l'utilisateur calculées à l'étape précédente et la structure du graphe utilisateur. Le résultat donne une représentation de l'ensemble du graphe désigné par z_G. L'apprentissage de la représentation des nœuds individuels, désignée par z_{u_i}, est également effectué à titre d'étape intermédiaire. Nous proposons dans cet article deux stratégies principales pour représenter l'ensemble du graphe pour les besoins de la détection de la controverse. Ces stratégies reposent sur des représentations hiérarchiques des graphes, des réseaux convolutifs et des représentations de graphes basées sur le mécanisme d'attention.

4.3.4.1 Stratégie basée sur l'apprentissage de représentation graphique hiérarchique.

Les GNN classiques pour l'encodage des graphes sont connus pour être plats. Ils ne propagent les informations que par les arêtes avant d'utiliser une simple couche d'agrégation et de regroupement (telle que les fonctions MAX ou MEAN) de toutes les représentations de nœuds résultantes. La stratégie que nous proposons dans cette section exploite la structure hiérarchique pouvant exister dans la structure du graphe de l'utilisateur. Ainsi, dans l'ensemble du processus d'encodage des graphes, les informations relatives aux graphes sont agrégées sur les arêtes de manière itérative et hiérarchique.

À chaque itération, les nœuds utilisateurs qui interagissent les uns avec les autres sont regroupés. Ce faisant, nous accordons plus d'importance aux informations d'interaction entre les utilisateurs. Cela pourrait aider le modèle à comprendre certains comportements utilisateurs, et à capturer des caractéristiques importantes pour classifier les posts comme controversé ou non.

Nous nous appuyons sur l'approche DIFFPOOL [Yin+18] qui encode l'ensemble du graphe en empilant plusieurs couches de regroupement ("Pooling" en anglais), avec une dernière couche ne représentant un graphe d'un seul nœud. Chaque couche de regroupement est composée de deux GNN distincts : l'un, appelé GNN_{embed}, apprend les nouvelles représentations de nœuds d'utilisateurs H, et l'autre, appelé $\text{GNN}_{pooling}$, apprend une matrice d'affectation S indiquant quels nœuds utilisateurs sont affectés à quel regroupement. La matrice S est utilisée ensuite pour réduire le graphe. Comme le montre la figure 4.3, le fonctionnement de la couche de mise en commun au niveau (l) est décrit comme suit :

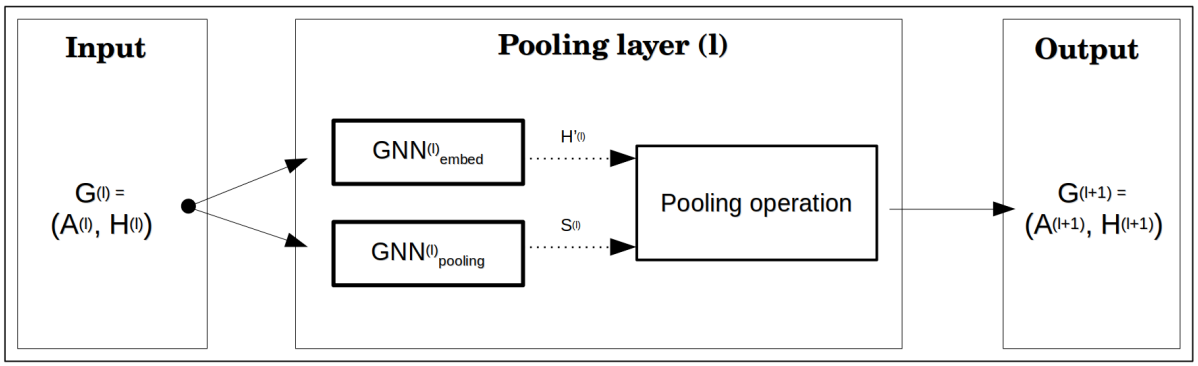

Figure 4.3 – *Architecture de la couche de regroupement basée sur DIFFPOOL. $A^{(l)}$ et $H^{(l)}$ représentent respectivement les matrices d'adjacence et de caractéristiques du graphe d'entrée à la couche (l). Les blocs GNN_{embed} et $\text{GNN}_{pooling}$ sont les deux blocs GNN utilisés pour calculer respectivement les nouvelles représentations des nœuds $H'^{(l)}$ et la matrice d'affectation $S^{(l)}$. Le bloc d'opération de mise en commun convertit le graphe d'entrée $(A^{(l)}, H^{(l)})$ en un nouveau graphe plus restreint $(A^{(l+1)}, H^{(l+1)})$.*

1. *Génération de la représentation des nœuds*. Nous appliquons d'abord la méthode $\text{GNN}^{(l)}_{embed}$ au graphe obtenu à la couche (l) représenté par sa matrice d'adjacence $A^{(l)}$, et sa matrice de caractéristiques des nœuds $H^{(l)}$. Comme décrit dans l'équation 4.2, le résultat est une représentation intermédiaire de nœuds $H'^{(l)} \in \mathbb{R}^{m \times d'}$, avec m le nombre de nœuds du graphe initial de la couche, et d' la dimension du vecteur des nouvelles caractéristiques.

$$H'^{(l)} = \text{GNN}^{(l)}_{embed}\left(A^{(l)}, H^{(l)}\right) \tag{4.2}$$

2. *Apprentissage des matrices d'affectation*. Nous utilisons ensuite le bloc $\text{GNN}^{(l)}_{pooling}$ pour apprendre une nouvelle matrice d'affectation $S^{(l)}$ à la couche (l), indiquant quels nœuds du graphe seront regroupés dans quels clusters. Ces clusters représenteront nos nouveaux nœuds dans le graphe restreint construit à la sortie de la couche (l). L'affectation matricielle est représentée par l'équation 4.3 :

$$S^{(l)} = \text{GNN}^{(l)}_{pooling}\left(A^{(l)}, H^{(l)}\right) \tag{4.3}$$

3. *Regroupement des nœuds, création du graphe restreint*. Nous regroupons enfin les nœuds appartenant au même groupe et leurs caractéristiques de $H'^{(l)}$ en utilisant

la matrice d'affectation $S^{(l)}$ pour produire le nouveau graphe restreint, représenté par sa matrice d'adjacence $A^{(l+1)}$ et sa matrice de caractéristiques $H^{(l+1)}$. Cette opération de regroupement s'effectue en suivant les deux équations 4.4 et 4.5 suivantes :

$$A^{(l+1)} = S^{(l)^T} A^{(l)} S^{(l)} \tag{4.4}$$

$$H^{(l+1)} = S^{(l)^T} H'^{(l)} \tag{4.5}$$

La figure 4.4 montre un exemple de graphe représenté à travers les différentes étapes d'une couche dans notre stratégie basée sur l'apprentissage de la représentation hiérarchique des graphes. Le nouveau graphe restreint en sortie de chaque couche représente le graphe d'entrée pour la couche suivante, où la représentation des nœuds a été calculée à l'aide des matrices H et S de la couche actuelle.

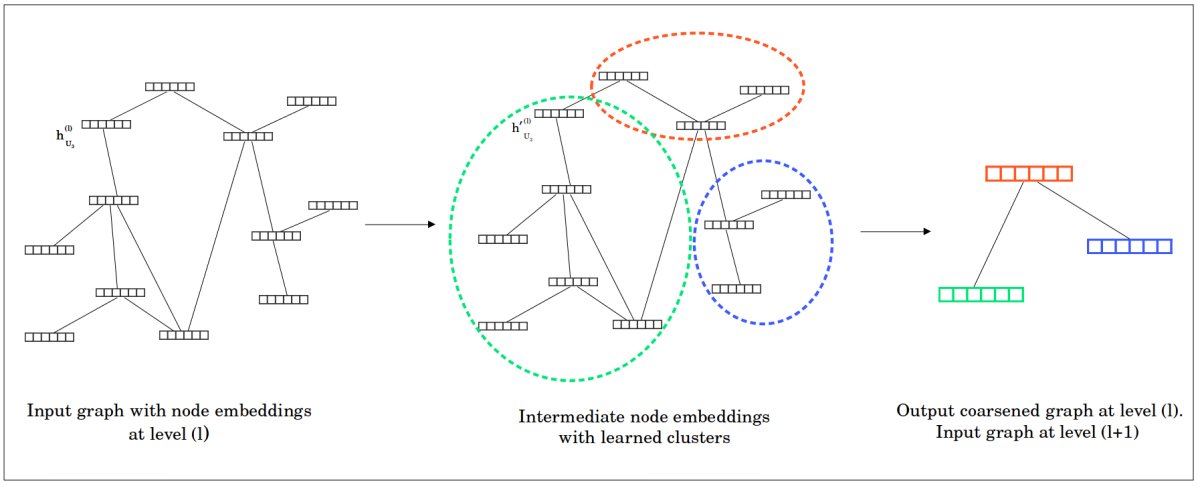

Figure 4.4 – *Exemple sur un graphe donné du fonctionnement d'une couche de notre approche, en utilisant une agrégation des nœuds par regroupement.*

On remarque que le nombre de nœuds diminue à chaque nouvelle couche (l). À la première couche (l=0), A^0 et H^0 correspondent respectivement à la matrice d'adjacence A et à la matrice de caractéristiques X du graphe utilisateur initial. La dernière couche L retourne un graphe d'un seul nœud, correspondant alors au vecteur de représentation final z_G de notre graphe.

Dans notre travail, un seul type de GNN est utilisé pour nos blocs $\text{GNN}^{(l)}_{embed}$ et $\text{GNN}^{(l)}_{pooling}$: Les réseaux de neurones graphiques à convolutions GCN [KW17] ("Graph Convolutional Networks" en anglais), que nous présentons.

GCN. Les GCNs [KW17] sont des modèles reconnus d'apprentissage pour les réseaux de neurones graphiques, et ont atteint des performances élevées dans plusieurs jeux de données de référence, pouvant représenter avec précision le voisinage local d'un nœud dans un espace de dimension d. Les GCNs, basés sur la théorie spectrale des graphes, agrègent la représentation des voisins avec elle-même en utilisant la matrice d'adjacence et de degré (respectivement A et D) de notre graphe utilisateur G afin d'obtenir une représentation latente de chacun de ses nœuds. Nous appliquons ici une transformation linéaire pour apprendre une matrice de poids $W^{(l')}$ à chaque couche

l'. Nous mettons à jour la représentation de tous les nœuds à l'aide de la règle de propagation montrée dans l'équation 4.6 suivante :

$$H^{(l'+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l')}W^{(l')}) \tag{4.6}$$

Où $\tilde{A}$ représente la matrice d'adjacence et $\tilde{D}$ la matrice de degré additionnée à la matrice d'identité (elle-même additionnée à la matrice identité, afin de prendre en compte sa propre ancienne représentation). $H^{(l'+1)}$ correspond à la nouvelle représentation des nœuds à la sortie de la couche l, et $H^{(l')}$ la représentation en entrée (au premier niveau, $H^{(0)} = X$). σ représente une fonction d'activation, ($ReLU$ dans notre cas, sauf pour la dernière couche où une fonction $SOTFMAX$ est utilisée).

Dans nos travaux, nous appliquons un GCN à deux couches, afin d'obtenir un bon équilibre entre les informations locales et globales du graphe dans notre représentation. $H^{(2)}$ sera donc notre représentation matricielle finale des nœuds, où $H^{(2)} \in \mathbb{R}^{N\times d'}$, d' correspondant à la dimension finale des caractéristiques de notre nœud unique, et donc de notre graphe.

4.3.4.2 Stratégie de regroupement des utilisateurs basée sur le mécanisme d'attention

Cette stratégie repose sur un calcul de la représentation des nœuds basée sur le mécanisme d'attention [Vas+17], permettant à chaque nœud d'utilisateur de traiter ces voisins en fonction de leur importance lors du processus de calcul de la nouvelle représentation d'un nœud, en fonction de la structure et des caractéristiques du graphe. Une fois que la représentation des nœuds est générée, elle est agrégée afin de produire un seul vecteur, faisant alors office de vecteur de représentation du graphe.

Le mécanisme d'attention (GAT) avec préservation de la cardinalité [ZX20] est utilisé pour différencier les voisins de l'utilisateur en leur attribuant des scores différents. Ce mécanisme d'attention est similaire à celui utilisé dans les blocs "Transformers" exploité par les modèles BERT [Dev+19] pour des tâches de TAL.

Soit $\tilde{\mathcal{N}}_{(u_i)}$ un ensemble de voisins de premier ordre du nœud u_i, y compris u_i lui-même. Cette deuxième stratégie est décrite comme suit :

1. *Score d'attention des voisins*. Pour chaque nœud $u_i \in \mathcal{U}$, et chaque utilisateur voisin $u_j \in \tilde{\mathcal{N}}_{(u_i)}$, nous calculons d'abord le score d'attention $e_{u_iu_j}$ en utilisant une fonction d'attention a sur les caractéristiques transformées représentées par la matrice $W^{(l)}$ de la couche actuelle (l) pour les deux nœuds, comme décrit dans l'équation 4.7 :

$$e^{(l)}_{u_iu_j} = a\left(\mathbf{W}^{(l)}h^{(l)}_{u_i}, \mathbf{W}^{(l)}h^{(l)}_{u_j}\right) \tag{4.7}$$

2. *Normalisation des scores d'attention*. Nous normalisons ensuite les scores à l'aide d'une fonction $SOTFMAX$ afin d'obtenir une distribution de probabilité pour

chaque score, comme le montre l'équation 4.8 :

$$\alpha_{u_i u_j}^{(l)} = softmax(e_{u_i u_j}^{(l)}) = \frac{\exp(e_{u_i u_j}^{(l)})}{\sum_{u_k \in \tilde{\mathcal{N}}_{(u_i)}} \exp(e_{u_i u_k}^{(l)})} \tag{4.8}$$

3. *Représentation des nœuds utilisateurs*. Les scores normalisés sont ensuite utilisés pour calculer la nouvelle représentation du nœud utilisateur $h_{u_i}^{(l+1)}$:

$$h_{u_i}^{(l+1)} = \sigma\left(\sum_{u_j \in \tilde{\mathcal{N}}_{(u_i)}} \alpha_{u_i u_j}^{(l)} \mathbf{W}^{(l)} h_{u_j}^{(l)}\right) \tag{4.9}$$

σ représentant une fonction d'activation non linéaire, et $h_{u_i}^{(l+1)} \in \mathbb{R}^{n \times d}$ la sortie de notre couche avec d sa dimension. À noter que l'ajout de la préservation de la cardinalité permet dans l'équation 4.9 de mettre à l'échelle le résultat avant l'utilisation de la fonction d'activation σ. Il est aussi possible d'appliquer un mécanisme d'attention à plusieurs têtes à chaque couche pour stabiliser le processus, en utilisant exactement le même processus K fois $(\alpha_{u_i u_j}^{k,(l)})W^{k,(l)}$, puis en concaténant les K représentations produites pour obtenir la représentation finale du nœud $h_{u_i}^{(l+1)}$ à la sortie de la couche l. La représentation finale du nœud h_{u_i} correspond à la sortie de la dernière couche $h_{u_i}^{(L)}$.

4. *Représentation du graphe*. Enfin, nous calculons la représentation finale du graphe z_G en appliquant une fonction $READOUT$. Une simple fonction d'agrégation au niveau du graphe est utilisée : nous concaténons, pour chaque nœud, ces représentations à chaque couche d'itération pour avoir une représentation globale, puis nous sommons toutes les représentations des nœuds afin de n'avoir plus qu'un vecteur représentant le graphe. Ce processus est représenté par l'équation 4.10.

$$z_G = \Big\|_{l=0}^{(L)} \left(READOUT\left(\{ h_{u_i}^{(l)} \Big\| u_i \in U \} \right) \right) \tag{4.10}$$

Cette méthode basée sur le mécanisme d'attention présente de multiples avantages, en plus des résultats élevés obtenus dans de nombreuses tâches de classification de graphes de référence. Son fonctionnement peut être utile dans le cadre de l'explicabilité des modèles. Il permet aussi d'utiliser un nombre fixe de paramètres (selon le nombre de voisins) indépendamment de la taille du graphe. Elle présente également des capacités inductives, permettant ainsi au modèle de faire des prédictions sur des graphes jamais vus.

4.3.5 Classification du graphe

Enfin, l'étape de classification du graphe vise à classer le message, représenté par sa nouvelle représentation du graphe z_G, comme controversé ou non controversé. Pour ce faire, nous nous appuyons simplement sur un classifieur perceptron multicouches [2],

2. Type de réseau de neurones comportant plusieurs couches de neurones interconnectés.

avec le vecteur z_G en entrée. Une fonction activation $SOTFMAX$ sur la couche de sortie de dimension deux est utilisée, afin de retourner la probabilité que le post soit controversé ou non.

4.4 Évaluation des expérimentations

Après avoir implémenté notre approche, nous avons mené des expériences sur le réseau social Reddit. L'ensemble de données comprend six jeux de données différents, correspondant à six subreddits, contenant plusieurs posts labellisés controversés ou non. Le jeu de données est présenté de manière détaillée dans la section 2.4.1.

4.4.1 Préparation des données

Nous commençons par séparer nos données en six jeux de données en fonction des subreddits. Le tableau 2.2, présenté dans la section 2.4.1, regroupe les statistiques des différents ensembles de données. Pour chaque subreddit s, nous créons un ensemble de graphes d'utilisateurs $\mathcal{G}_s$, un graphe par post. Chaque ensemble présente un minium de 1 000 posts. Nous définissons ensuite $\mathcal{G}_{s,train}$ et $\mathcal{G}_{s,val}$ respectivement comme nos ensembles de graphes d'entrainements et de validation. Ces deux ensembles sont tous équilibrés entre les nombres de posts controversés et non controversés.

Compte tenu de tous les aspects et des expériences menées, nous évaluons notre approche toujours sur le même ensemble de validation. Comme l'ensemble de validation est équilibré, la mesure de précision est utilisée pour comparer les performances des approches avec celles de la littérature, que l'on considère comme notre base de référence.

Concernant l'apprentissage de la représentation des textes de l'utilisateur, notre modèle de langage est entraîné séparément (sur le même ensemble $\mathcal{G}_{s,train}$). Le modèle GNN, utilisé afin d'apprendre une nouvelle représentation du graphe, est entraîné par la suite.

4.4.2 Base de référence

Nous avons comparé notre approche avec des travaux de la littérature travaillant sur la tâche de détection de la controverse. Ces travaux ont été sélectionnés pour deux raisons principales. Premièrement, le même objectif est partagé, à savoir la détection des posts controversés, et couvrent les mêmes aspects textuels et structurels que nous avons abordés dans nos travaux. Deuxièmement, les expériences menées dans ces travaux sont basées sur les mêmes jeux de données Reddit.

- (**POST** (**Text+Time**)) [HL19]. Cette méthode se concentre uniquement sur le contenu des messages. Elle utilise la modélisation linguistique basée sur BERT [Dev+19] et des caractéristiques supplémentaires basées sur l'horodatage du message.
- (**C-{Text_Rate_Tree} + Post**) [HL19]. Cette méthode se base sur un simple classifieur binaire. Les vecteurs de représentation textuels d'un post sont combinés

avec des caractéristiques structurelles de l'arbre de commentaires (profondeur de l'arbre, nombre moyen de commentaires pour chaque commentaire, proportion de commentaires du premier niveau ayant au moins une réponse, etc.). Nous comparons les messages et les commentaires au cours de la première heure et des trois premières heures.

- (**DTPC-GCN**) [Zho+20]. Cette méthode est basée sur un réseau de neurones à convolutions graphiques, associant le subreddit à l'entrainement du modèle. Les messages controversés sont identifiés à l'aide d'un modèle GCN et donc de l'apprentissage de caractéristiques structurelles, mais aussi des caractéristiques textuelles. Le modèle est expérimenté pour la détection inter-sujets, où tous les messages sont rassemblés dans un seul ensemble de données.

4.4.3 Première expérimentation : Détection des posts controversés sur la base des informations structurelles

Nous implémentons dans l'approche de détection de la controverse basée sur les GNN en utilisant la librairie PyTorch. L'apprentissage de la représentation graphique hiérarchique se base sur les préceptes présentés par les méthodes DIFFPOOL [Yin+18] et GCN [KW17]. Nous appelons cette méthode **HRL-GCN** ("Hierarchical Representation Learning based on GCN" en anglais). Nous testons cette stratégie en utilisant respectivement une et deux couches de regroupement dans le réseau. Pour chaque couche de regroupement, un GCN utilisant trois couches est utilisé. Nous nous appuyons sur les mêmes fonctions de perte et d'optimisation que celles utilisées dans les expérimentations faites par Ying et al. [Yin+18] pour la méthode DIFFPOOL.

L'apprentissage de la représentation graphique se base sur le mécanisme d'attention présenté par Velickovic et al. [Vel+18] dans la méthode GAT, en se basant sur l'implémentation de la méthode GAT-GC [ZX20], et utilise le même paramétrage du réseau de neurones. Nous appelons cette méthode **ARL-GAT** ("Attention Representation Learning based on GAT" en anglais). Nous testons cette stratégie en utilisant deux agrégateurs de nœuds différents pour calculer la représentation du graphe, à savoir les fonctions de regroupement $MEAN$ et SUM. Concernant les hyper-paramètres, les deux stratégies basées sur les GNN sont entraînées avec un taux d'apprentissage fixe de 0,01, d'une taille de lot (ou "batch") de 32, le tout durant 100 étapes d'apprentissage (ou "epoch").

Les premières expériences sont réalisées sans utiliser la représentation textuelle de chaque utilisateur afin de souligner l'importance de l'interaction structurelle entre les utilisateurs dans les discussions controversées. La représentation textuelle de chaque utilisateur est simplement remplacée par le degré de son nœud, c'est-à-dire le nombre d'arêtes lui étant incidents dans le graphe. Le tableau 4.1 présente les résultats de cette expérience, mesurée par la précision de l'approche sur les jeux de validation. Les quatre premières lignes correspondent aux approches de notre base de référence, et les quatre dernières lignes aux résultats de nos expériences. La précision des différentes méthodes est indiquée pour chaque subreddit (AM, AW, FN, LS, PF, RS) décrit dans la section 2.4.1, à l'exception de la méthode DTPC-GCN pour laquelle la précision est liée à l'ensemble du jeu de données.

Table 4.1 – *Comparaison de la précision de nos approches basées sur les GNN avec notre base de référence sur la détection des posts controversés. La performance est évaluée en utilisant la précision sur l'ensemble de validation. Les valeurs en gras représentent les meilleurs scores parmi toutes les méthodes (y compris celles de références), tandis que les valeurs soulignées représentent les meilleurs scores parmi les méthodes que nous avons proposées.*

	AM	AW	FN	LS	PF	RS
POST (Text+Time)	68,1	65,4	65,5	66,2	66,5	69,3
DTPC-GCN			67,6			
POST + C-{Text_Rate_Tree} $<$ 1 hour	71,1	70	68,1	67,9	66,1	65,5
POST + C-{Text_Rate_Tree} $<$ 3 hours	**74,3**	72,3	70,5	**71,8**	**69,3**	**67,8**
ARL-GAT (mean-aggr)	65,7	69,2	**72,4**	58,4	53,7	62,9
ARL-GAT (sum-aggr)	67,5	71	72,2	67	63,7	51,8
HRL-GCN (pool=2)	69	72,2	71,7	68,3	65,7	63,6
HRL-GCN (pool=1)	69,6	**74,6**	72,2	67,9	68,2	66,7

Comme le montre le tableau 4.1, notre approche hiérarchique HRL-GCN obtient les meilleurs résultats sur plusieurs jeux de données, en utilisant une seule couche de regroupement. Notre approche basée sur l'attention ARL-GAT atteint des performances moyennes moins importantes, que ce soit avec l'agrégateur sum ou mean. HRL-GCN surpasse la méthode DTPC-GCN [Zho+20] et la méthode hybride proposée par Hessel et Lee [HL19] avec les commentaires émis lors de la première heure après la publication du post pour 5 des 6 jeux de données. La méthode que nous proposons (HRL-GCN, pool=1) obtient des résultats proches de l'état de l'art sur plusieurs jeux de données, allant même jusqu'à une précision de $74,6$ dans le jeu de données aw, surpassant les résultats de la méthode C-{Text_Rate_Tree} + Post comprenant des commentaires de plus de trois heures après la publication du post. Le jeu de données am étant le plus grand, cela pourrait signifier que notre approche se généralise mieux lorsque les données sont abondantes. Comme expliqué par Hessel et Lee [HL19], le peu de commentaires disponibles sous chaque post explique en partie la difficulté à obtenir de meilleurs résultats. En effet, lorsque le subreddit *AskMen* (am) a en moyenne 10 commentaires après 45 minutes, *Relationships* (rs) n'en a même pas 10 après 3 heures.

Le tableau 4.1 montre aussi que notre approche basée sur l'attention ARL-GAT, combinée avec l'agrégateur $mean$, est performante dans les trois premiers jeux de données, battant notre meilleure méthode de référence sur fn, avec une précision de $72,4$ sur l'ensemble de validation. En revanche, elle est moins performante sur les trois autres jeux de données, avec une précision de $53,7$ sur pf. Les jeux pf et rs ont déjà montré des résultats faibles avec les modèles de référence, signifiant déjà la complexité de la tâche. Cela pourrait également s'expliquer par le fait que ces 3 subreddits ont le nombre moyen de commentaires le plus faible (comme le montre le tableau 2.2), et que chaque nœud utilisateur a donc moins de voisins. Les scores d'attention sont en fait moins utiles dans ces cas. Un degré moyen plus élevé des nœuds pourrait conduire à de meilleures performances.

4.4.4 Seconde expérimentation : Détection des posts controversés basée l'enrichissement du graphe par le contenu textuel.

Nous avons mené une deuxième session d'expérimentations afin d'étudier l'impact de l'ajout de caractéristiques du contenu textuels de chaque utilisateur à nos différentes architectures. Au lieu d'examiner toutes les options de nos différentes architectures basé sur les GNN comme dans la section 4.4.3, nous n'avons pris en compte que la stratégie de notre représentation hiérarchique HRL-GCN, avec une seule couche de regroupement, car elle présente les meilleurs résultats en termes de précision dans le tableau 4.1.

Les caractéristiques textuelles des commentaires et des posts sont extraites à l'aide de différents modèles de langage basés sur BERT, et sont agrégées par utilisateur pour être utilisées comme caractéristiques initiales de nos nœuds utilisateurs dans nos graphes. Nous testons différents modèles pour extraire ces caractéristiques :

- Modèle **PT**. Ce modèle n'utilise que les caractéristiques pré-entraînées du modèle BERT pour obtenir la représentation d'un message. La dernière couche (768 dimensions) est restituée afin de servir de vecteur de représentation du message.
- Modèle **FT_ITSELF**. Dans ce modèle, nous affinons un modèle BERT [Sun+19] en utilisant les commentaires et les posts de notre ensemble d'entraînement $\mathcal{G}_{s,train}$, avec une couche supplémentaire de 64 neurones, en plus d'une couche de classification. Nous étiquetons chaque commentaire selon son post respectif, avec le label controversé ou non. Il convient de noter que chaque subreddit est affiné séparément, et donc à l'aide de modèles différents, car nous supposons que les différentes communautés s'expriment différemment et que les messages peuvent être interprétés différemment.
- Modèle **FT_SENTIMENT**. Nous affinons un modèle BERT en utilisant l'analyse de sentiments avec un autre ensemble de données Reddit de commentaires [3], étiquetés comme négatifs, positifs ou neutres. En effet, nous supposons ici que les sentiments peuvent donner un aperçu du comportement des utilisateurs dans des discussions controversés.
- Modèle **PT_LSTM**. Les modèles basés sur BERT ne peuvent accepter plus de 512 tokens au maximum. Pour contrer le fait que les messages sur Reddit ne sont pas limités et peuvent contenir des textes longs, nous créons des lots de taille fixe (200 mots avec un chevauchement de 50 mots) pour chaque message, et nous extrayons leur représentation d'un modèle BERT pré-entraîné. Sur la base de l'idée présentée par PAPPAGARI et al. [Pap+19], nous formons ensuite un modèle bi-LSTM avec la représentation vectorielle de chaque lot, en suivant le même principe que dans FT_ITSELF pour obtenir finalement une représentation globale du message à la fin.

Dans chacun de ces cas, nous utilisons la version "base-bert-uncased" (avec son tokenizer correspondant), avec ces 12 couches "Transformers" et 110 millions de paramètres. Pour des raisons de temps et de mémoire, nous n'utilisons un maximum de 256

3. Ces données proviennent des serveurs de kaggle.com

Table 4.2 – *Performance de notre meilleure approche basée sur les GNN enrichis avec différentes représentations textuelles des utilisateurs comme caractéristiques initiales de nœuds. Les valeurs en gras représentent les meilleures précisions parmi toutes les méthodes (y compris celles de notre base de référence), tandis que les valeurs soulignées représentent les meilleures précisions parmi les méthodes que nous avons proposées.*

	AM	AW	FN	LS	PF	RS
HRL-GCN (pool=1)	69,6	74,6	72,2	67,9	68,2	66,7
+ FT_SENTIMENT	69,1	72,9	70,5	68,6	66,7	64
+ FT_ITSELF	67,3	73,9	71,8	68,3	70,6	63,8
+ PT	70,8	73,7	71	65,4	0,6	64,7
+ PT_LSTM	69,6	**74,8**	72,3	68,9	**70,7**	65,1

tokens maximum pour chaque message (au lieu de 512, le tableau 2.2 montrant qu'en moyenne moins de 3% des messages sont représentés par plus de 256 tokens). Pour affiner ces modèles, nous avons utilisé les mêmes hyperparamètres que ceux présentés par Sun et al. [Sun+19]. Le tableau 4.2 montre les nouveaux scores de précision obtenus en incorporant les caractéristiques textuelles dans notre stratégie HRL-GCN.

Pour cinq de nos six jeux de données, l'ajout de caractéristiques textuelles améliore, même légèrement, les résultats de la détection des posts controversés. Notre stratégie HRL-GCN, fusionnée avec les caractéristiques du modèle BERT pré-entraîné combiné à un bi-LSTM (PT_LSTM), obtient de meilleurs résultats lors de l'utilisation des ensembles de données AW, FN, LS, PF, avec respectivement $74,8$, $72,3$, $68,9$ et $70,7$ de précision. Cependant, la combinaison n'améliore pas les performances sur AM, où l'ajout des seules caractéristiques BERT pré-entraînées (PT) améliore la précision du modèle en atteignant un score de $70,8$. L'ajout des caractéristiques textuelles représentant les sentiments des messages (FT_SENTIMENT) nous permet également d'augmenter la précision de $67,9$ à $68,6$ sur LS. Cependant, les performances ne sont pas aussi bonnes que celles de notre modèle PT_LSTM, accentuant le fait que les sentiments seuls ne sont pas représentatifs de la controverse, et ne peuvent définir un message appartenant à un post controversé. La complexité des données et le fait que le nombre de nœuds par post soient trop petits par rapport au nombre de caractéristiques par nœud (768 en utilisant le modèle PT) pourraient également expliquer pourquoi les modèles sur-apprennent sur certains jeux de données, et n'améliorent donc pas les résultats en termes de précision. La réduction de la dimension appliquée dans la dernière couche de PT_LSTM permet en partie de réduire ce problème.

La figure 4.5 montre aussi l'importance de la disponibilité des commentaires à travers le temps. Elle présente l'évolution des résultats de précision en fonction du temps (minutes) pour trois jeux de données en utilisant notre meilleure stratégie HRL-GCN combinée avec les caractéristiques textuelles du modèle PT. Il apparaît clairement que plus les commentaires sont disponibles, plus il est facile de détecter qu'un post est controversé ou non. De plus, concernant les statistiques des jeux de données présentés dans le tableau 2.2, cela pourrait également expliquer pourquoi notre méthode obtient des résultats inférieurs sur les subreddits LT et RS par rapport

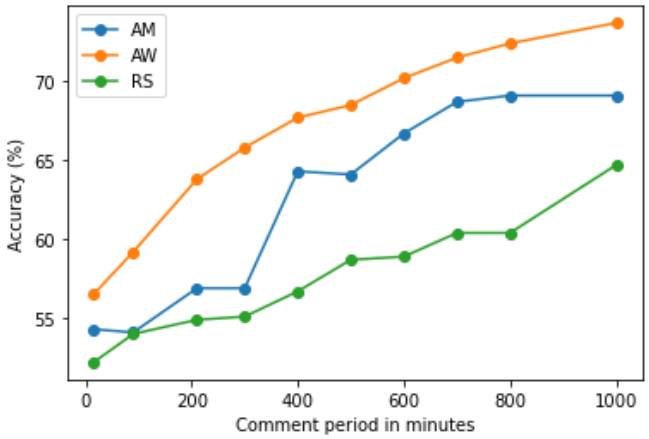

FIGURE 4.5 – *Impact de la disponibilité des commentaires dans le temps sur les performances de la détection de posts controversés, en se basant sur notre approche HRL-GCN.*

aux autres jeux de données.

4.5 Application de notre approche sur des graphes polarisés : Le cas de Twitter

Chaque réseau social a des fonctionnalités différentes et induisent des comportements spécifiques chez les utilisateurs. C'est le cas notamment du Retweet sur Twitter. Le Retweet consiste à partager le contenu exact d'un tweet original d'un autre utilisateur. Il est souvent considéré comme une sorte d'approbation de ce contenu. Dans le cas de la controverse, cette action est très intéressante, car elle permet notamment à un utilisateur de partager l'opinion ou le point de vue d'un autre. De ce fait, les utilisateurs peuvent être polarisés autour de communautés ayant des points de vue communs. Ainsi, l'analyse et la détection de la controverse des graphes utilisateurs de Retweet consiste à étudier la controverse d'un point de vue de la polarisation des communautés utilisateurs. Cela a notamment été mis en avant lors des travaux de GARIMELLA et al. [Gar+18a].

Dans nos travaux, nous allons voir si l'approche présentée dans la section 4.3 est généralisable pour tout type de graphes et de données. Ainsi, nous allons étudier les performances de détection de la controverse sur des sujets Twitter.

4.5.1 Méthodologie

Concernant la méthodologie, seul le traitement des données et la création du graphe change. Dans la section 4.3.2, nous présentions la création d'un graphe utilisateur à partir des échanges entre utilisateurs (sous forme d'arbre de commentaire). Pour ce qui est des données Twitter, plusieurs types de graphes utilisateurs peuvent être créés à partir des données fournies par l'API, tels qu'un graphe de suivi ("follow" en anglais), un graphe de Retweet, un graphe de citation, ou même un graphe hybride combinant plusieurs types de liens entre utilisateurs. Afin de nous rapprocher des expériences

menées sur Reddit, nous avons décidé de nous en tenir à la création de graphes de retweet (RT). Pour rester cohérent avec la notation Reddit, nous construisons, pour chaque sujet p, un graphe non dirigé où un nœud utilisateur u_i est lié à un autre utilisateur u_j si l'un d'eux a retweeté l'autre au moins une fois. Chaque utilisateur est représenté par les tweets originaux qu'il a posté. Les graphes de retweets ont des propriétés intéressantes, telles qu'une bonne représentation de l'interaction des utilisateurs et de la séparation des communautés.

4.5.2 Évaluation des Expérimentations

4.5.2.1 Préparation des données

Le jeu de données utilisé lors de nos expérimentations est introduit dans la section 2.4.2 et contient 15 sujets au total (neuf controversés et six non controversés). Les statistiques de ces jeux de données sont présentées dans la table 2.3. À partir de ces données, nous créons des graphes de Retweet utilisateurs, représentant, pour chaque sujet, un graphe dans lequel les utilisateurs (nœuds) sont liés les uns aux autres si l'un a retweeté l'autre au moins une fois (arête). Nous supposons que cette représentation graphique nous aidera à représenter des groupements de communautés sur des graphes controversés, comme expliqué par Garimella et al. [Gar+18a]. En utilisant la même méthode dans [Gar+18a] avec metis [4], les graphes sont partitionnés en deux communautés, pour voir si le nombre de tweets récupérés pour chaque sujet reste suffisant pour capturer l'information structurelle.

Notre principale préoccupation concernant ces expériences sur les données Twitter est double : le volume des graphes de retweet avec un nombre très important de nœuds, et le faible nombre de graphes au total que nous pouvons obtenir (seulement 15 sujets disponibles). Pour résoudre ces deux problèmes, nous nous sommes appuyés sur des techniques d'échantillonnage de graphes et d'augmentation des données.

Méthodes d'échantillonnage de graphes. Les 15 graphes rassemblés dans l'ensemble $\mathcal{G}_s$ sont malheureusement trop volumineux pour être traités par nos modèles, même avec une petite taille de lot. À chaque étape du processus d'apprentissage, le modèle prend l'ensemble du graphe en entrée. Sur les différents jeux de données Reddit, le nombre moyen d'utilisateurs par post est de 65. Avec une matrice d'adjacence de taille n^2 (avec n le nombre d'utilisateurs), cela reste acceptable en termes de capacité pour nos machines. Cependant, nous travaillons sur une échelle différente sur notre ensemble de données Twitter, avec un nombre moyen d'utilisateurs d'environ 52 000 sur nos sujets. Par conséquent, la RAM de nos machines ne peut pas le supporter, même en utilisant des lots d'un graphe, et nous nous retrouvons rapidement à court de mémoire. Par conséquent, des techniques d'échantillonnage de graphes sont nécessaires pour réduire la taille de nos différents grands graphes.

Les méthodes d'échantillonnage de graphes visent à dériver un petit graphe similaire à partir du graphe original [HL13]. L'opération d'échantillonnage consiste à sélectionner un sous-ensemble de sommets et/ou d'arêtes du graphe original. Dans les expériences que nous avons menées, nous nous sommes appuyés sur trois méthodes

4. Bibliothèque utilisée pour le partitionnement des graphes en série et l'ordonnancement de matrices.

d'échantillonnage, respectivement ISRW, FF et TIES, afin de réduire la taille des larges graphes générés à partir de Twitter [5]. Nous présentons ces méthodes ci-dessous.

1. **ISRW** : Échantillonnage par marche aléatoire sur des sous-graphes induits ("Induced Subgraph Random Walk Sampling" en anglais). Il s'agit d'un d'échantillonnage de nœuds orienté, basé sur l'algorithme classique d'échantillonnage par marche aléatoire. L'algorithme commence par choisir au hasard un nœud de départ. Ensuite, il effectue une marche aléatoire en sélectionnant, au hasard, un nœud voisin sur le graphe. Il se poursuit jusqu'à ce que la taille de l'échantillon de nœuds requis soit atteinte. Pour éviter d'obtenir des nœuds avec des degrés inférieurs à ceux des nœuds d'origine, cette méthode ajoute une étape d'induction de graphe pour sélectionner des arêtes supplémentaires entre les nœuds échantillonnés dans le but de restaurer la connectivité et d'obtenir une meilleure similarité au niveau de la structure avec le graphe d'origine.

2. **FF** : Feu de forêt ("Forest Fire" en anglais). Il s'agit également d'une méthode d'échantillonnage basée sur les nœuds. Elle choisit au hasard un nœud de départ et commence par ce que l'on appelle "brûler" les arêtes sortantes et leurs nœuds correspondants. Si un lien est brûlé, le nœud à l'autre extrémité a la possibilité de brûler ses propres liens. Ce processus est répété de manière récursive pour chaque voisin brûlé jusqu'à ce qu'aucun nouveau nœud ne soit sélectionné. Ce processus est répété avec un nouveau nœud aléatoire comme nœud de départ, jusqu'à ce que la taille souhaitée de l'échantillon soit atteinte.

3. **TIES** : Échantillonnage des bords par induction totale. Contrairement aux deux méthodes ci-dessus, TIES est une méthode d'échantillonnage orientée autour des arêtes. Elle fonctionne de manière itérative, en choisissant une arête au hasard dans le graphe original et en ajoutant les deux nœuds à l'ensemble des nœuds échantillonnés à chaque itération. L'ajout de nœuds s'arrête lorsque la fraction cible φ de nœuds est collectée. Ensuite, l'algorithme passe à l'étape d'induction du graphe où il parcourt toutes les arêtes du graphe et forme le graphe induit en ajoutant toutes les arêtes dont les deux extrémités se trouvent déjà dans l'ensemble de nœuds échantillonné.

Les méthodes d'échantillonnage des graphes ayant des stratégies différentes, les graphes obtenus ne sont donc pas similaires. La figure 4.6 montre que les méthodes d'échantillonnage ISRW et TIES semblent présenter une meilleure similarité avec la structure des graphes controversé originaux (deux communautés sont bien représentées et séparées). Toutefois, TIES utilisant l'échantillonnage par les arêtes, il doit avoir accès à l'ensemble des arêtes. Dans le cas de certains graphes très larges, le nombre d'arêtes étant plus important par rapport au nombre de nœuds, la méthode TIES peut devenir moins efficace en termes de calcul lorsqu'il est appliqué à ces très grands graphes. Par conséquent, même si nous testons les deux options, nous faisons l'hypothèse préliminaire que si nous devons choisir entre ces deux méthodes, la méthode ISRW est plus appropriée. D'autre part, la méthode FF semble se concentrer sur la représentation d'autres aspects du graphe, ce qui rend aussi son analyse intéressante. la Figure 4.7 montre la même conclusion sur des graphes non controversés, montrant

5. https://github.com/Ashish7129/Graph_Sampling

une moins bonne séparation.

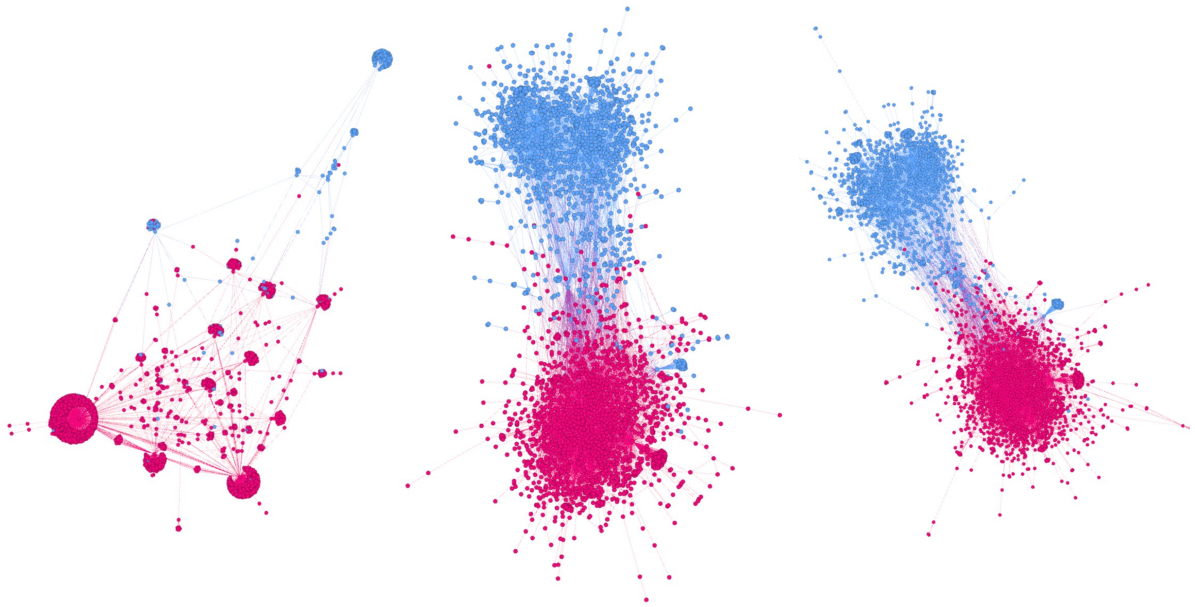

Figure 4.6 – *Différents graphes de Retweet échantillonnés pour le* ***sujet controversé "netanyahu"****. À gauche : graphe échantillonné avec FF. Au milieu : graphe échantillonné avec TIES. À droite : graphe échantillonné avec ISRW.*

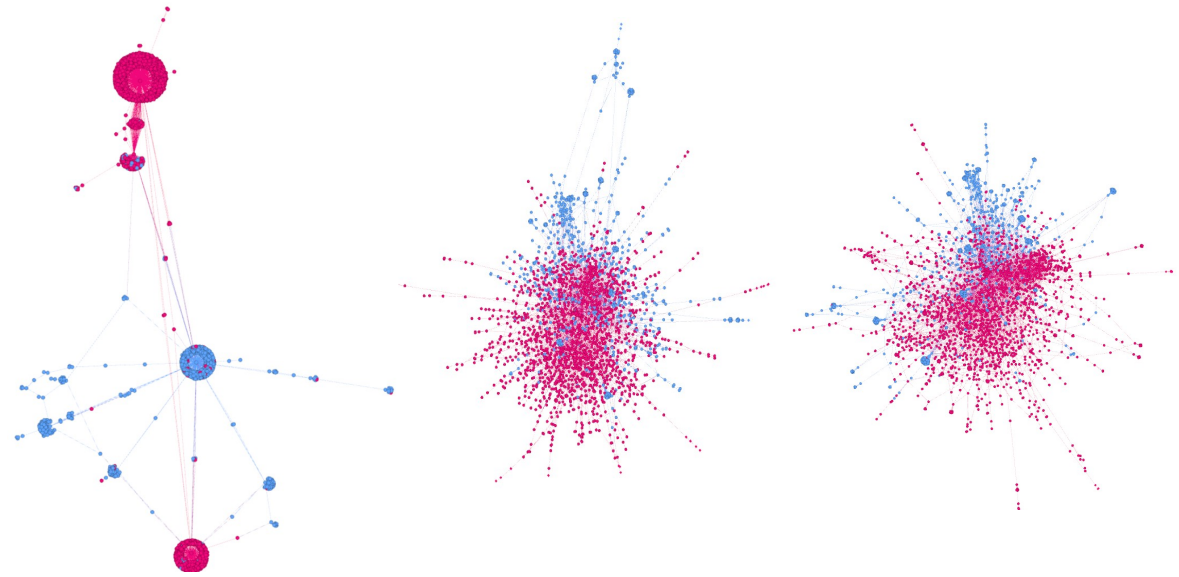

Figure 4.7 – *Différents graphes de Retweet échantillonnés pour le* ***sujet non controversé "sxsw"****. À gauche : graphe échantillonné avec FF. Au milieu : graphe échantillonné avec TIES. À droite : graphe échantillonné avec ISRW.*

Augmentation des données. Contrairement à Reddit, où il est facile de rassembler

de nombreux sujets autour d'un subreddit, il est compliqué de localiser et extraire les sujets et leurs tweets correspondant sur Twitter. L'échantillon d'entraînement étant trop petit, nous avons utilisé la méthode $leave - one - out$ pour diviser les données en 15 plis différents. Par conséquent, un seul graphe sera utilisé dans l'ensemble de test $\mathcal{G}_{i,val}$ à chaque fois, et les 14 autres seront utilisés comme échantillons lors de l'entraînement $\mathcal{G}_{i,train}$ à chaque itération i. La précision de notre modèle est ainsi calculée en faisant la moyenne des précisions calculée à partir des différents plis. Le graphe de test sera également réduit en utilisant respectivement les différentes méthodes d'échantillonnage ISRW, FF et TIES. Cela implique que nous formions 15 modèles différents pour chaque expérience. Pour pallier le nombre insuffisant de graphes disponibles dans l'ensemble $\mathcal{G}$, nous nous sommes appuyés sur des techniques d'augmentation de données en se basant sur les méthodes d'échantillonnage ISRW, FF et TIES présenté précédemment, afin d'augmenter l'ensemble d'entraînement. Nous avons créé 10 sous-graphes pour chaque sujet du jeu d'entraînement $\mathcal{G}_{i,train}$. Nous avons ensuite équilibré notre jeu d'entraînement en sélectionnant le même nombre de graphes échantillonnés dans les deux classes.

En ce qui concerne les caractéristiques textuelles, nous avons utilisé deux modèles basés sur le modèle BERT pour représenter les tweets en un vecteur dans un espace euclidien :

1. Modèle **PT**. Il s'agit du même modèle BERT pré-entraîné que nous avons utilisé dans la section 4.4.4 sur notre ensemble de données Reddit.
2. Modèle **BERTweet**. Il s'agit d'un modèle de langage spécialement pré-entraîné pour les tweets en anglais par Nguyen, Vu et Nguyen [NVN20]. BERTweet est entraîné sur la base de la procédure de pré-entraînement RoBERTa, une autre méthode construite sur BERT mais avec des hyperparamètres spécifiques et propres.

Nous avons mené des expérimentations sur des modèles avec et sans caractéristiques textuelles, comme nous l'avons fait avec les données Reddit, afin de voir si l'information textuelle à aussi un impact sur les performances de notre approche sur de larges graphes de Retweet. À noter que nous n'avons testé que le modèle ayant obtenu les meilleurs résultats (sans caractéristiques textuelles) sur l'ensemble de données Reddit, à savoir le modèle HRL-GCN (en utilisant une seule couche de regroupement).

4.5.2.2 Troisième expérimentation : Détection des posts controversés à partir des graphes de retweets

Nous avons mené des expériences en utilisant les trois méthodes d'échantillonnage présentées dans la section 4.5.2.1 afin de réduire la taille du graphe et augmenter le nombre de graphes échantillonnés. Concernant les méthodes de marche aléatoire (ISRW) et de feu de forêt (FF), nous fixons le nombre maximal de nœuds échantillonnés à 2 930, ce qui correspond à 1% du nombre moyen des nœuds des graphes entiers. Concernant la méthode d'échantillonnage des bords (TIES), nous avons fixé φ de façon à ce que le nombre maximal de nœuds échantillonnés corresponde à 1% du graphe entier. Nous avons testé notre approche avec et sans caractéristiques textuelles des nœuds sur notre approche HRL-GCN. Comme nous avons utilisé le leave-one-out

Table 4.3 – *Performances en termes de précision de notre méthode HRL-GCN sur la détection de sujets controversés sur Twitter. La colonne "Modèle de langage" représente le modèle utilisé pour calculer les caractéristiques textuelles de l'utilisateur. "Degré du nœud" correspond aux expériences sans caractéristiques textuelles. Les valeurs en gras représentent les meilleures précisions parmi les méthodes proposées.*

Méthode d'échantillonnage	**Modèle de langage**		
	Degré du nœud	PT	BERTweet
ISRW	0,5	0,47	0,47
TIES	0,53	0,53	0,5
FF	**0,67**	**0,67**	0,5

comme technique de validation du modèle, nous avons estimé la précision moyenne à partir de chaque pli.

D'après les résultats du tableau 4.3, l'échantillonnage par les méthodes ISRW et TIES n'obtiennent pas de bonnes performances, en comparaison avec celles de l'approche utilisée dans Garimella et al. [Gar+18a]. Au vu des graphes controversés émis par ces méthodes (figure 4.6), cela peut signifier que notre approche n'est pas performante lorsqu'elle tente de capturer les informations structurelles séparant les utilisateurs en communautés distinctes, se concentrant sur d'autres caractéristiques structurelles. De plus, le fait que nous n'ayons entraîné notre modèle que sur 14 graphes utilisateurs réels pourrait expliquer pourquoi le modèle n'est pas performant. Concernant la représentation textuelle des utilisateurs, les modèles pré-entraînés semblent ne pas réussir à capturer les positions des utilisateurs sur des sujets controversés au sein des communautés, en particulier avec BERTweet. Le bruit dans nos données pourrait également expliquer pourquoi il est difficile de capturer des vecteurs textuels représentatifs sur le sujet actuel. Cependant, le tableau 4.3 montre que l'utilisation de la méthode d'échantillonnage de feu de forêt (FF) pour réduire et augmenter les graphes permet d'obtenir de meilleures performances. La méthode d'échantillonnage de feu de forêt (FF) est basée sur la propagation du feu, ce qui pourrait correspondre à un sujet controversé sur une position ou une idée sur un graphe utilisateur. En d'autres termes, il capture mieux les informations structurelles de notre graphe de retweets. Nous obtenons une précision de $0,67$ avec et sans les caractéristiques textuelles du modèle PT. Les résultats sont proches de ceux obtenus lors des expériences précédentes sur les ensembles de données Reddit. Nous estimons que cela montre que notre approche contient le potentiel pour être généralisée à d'autres réseaux sociaux, à condition d'avoir de meilleures données d'un point de vue quantitative (plus de sujets) et qualitative (relatif au sujet, surtout dans le cas de Twitter).

4.6 Conclusion

Nous présentons dans ce chapitre une méthode afin de détecter automatiquement la controverse sur les médias sociaux. Nous avons considéré la détection de la controverse

comme un problème de classification, sous deux angles différents, correspondant à l'étude de deux réseaux sociaux aux caractéristiques différentes. Nous avons adopté des techniques récentes basées sur les GNN et les modèles de traitement du langage naturel basé sur l'apprentissage automatique pour combiner à la fois les informations structurelles (interactions entre les utilisateurs) et le contenu des messages des utilisateurs concernés. Les messages postés sur n'importe quel média social sont représentés sous la forme d'un graphe utilisateur.

La représentation des graphes est réalisée en incorporant les caractéristiques textuelles des utilisateurs, calculées à partir de leurs contenus sur des modèles de type BERT et LSTM. L'approche proposée prend en charge deux stratégies de détection de la controverse. La première stratégie exploite la structure hiérarchique pouvant exister dans les graphes utilisateurs. La seconde stratégie permet à chaque utilisateur de sélectionner ses voisins dans le processus de représentation des nœuds. L'évaluation expérimentale a été réalisée sur différents jeux de données collectées sur les plateformes Reddit et Twitter. Cette évaluation confirme que les informations structurelles sont utiles pour détecter la controverse, et montre clairement que les caractéristiques textuelles peuvent améliorer la précision dans certains cas sur Reddit, tandis que nos expériences sur Twitter suggèrent que les modèles classiques pré-entraînés ne conviennent pas forcément. L'affinement d'un modèle basé sur BERT sur une tâche spécifique pourrait être une perspective intéressante, comme l'analyse de sentiments ou la détection de prise de position des utilisateurs.

En termes de travaux futurs, nous aimerions examiner la pertinence d'autres techniques basées les GNN pour la détection de la controverse sur les médias sociaux. Par exemple, il pourrait être intéressant d'étudier l'impact, en termes de performances, de l'utilisation d'architectures GNN prenant en compte les propriétés des nœuds et des arêtes [SK17] sur des graphes hétérogènes, où les nœuds des utilisateurs pourraient être liés de différentes manières. Concernant Twitter, combiner notre approche sur les réseaux de neurones graphiques, avec l'idée de quantifier la controverse [Gar+18a] sur un graphe entier, pourrait également être une piste intéressante à explorer. En effet, nous pourrions nous concentrer uniquement sur l'apprentissage d'informations structurelles avec les GNN, et étudier quelles mesures pourraient aider à comprendre la représentation du graphe et à quantifier la controverse. Afin d'avoir une meilleure vue d'ensemble des différentes structures, un jeu de données plus complet de sujets controversés et non controversés est nécessaire et pourrait donner lieu à des améliorations notables des performances de notre approche.

Ce chapitre s'est donc concentré sur la détection de la controverse, en se focalisant sur les discussions entre utilisateurs dans un premier temps (sur Reddit), puis sur la polarisation de ces utilisateurs dans un deuxième temps (sur Twitter). Détecter la controverse de manière binaire, ne permet pas de capturer de nuance dans l'intensité de la controverse. Dans cette optique, nous présentons, dans le chapitre 5, une nouvelle approche, basée sur les utilisateurs et utilisant les GNN, afin de quantifier la controverse sur Twitter.

CHAPITRE 5

QUANTIFICATION DU POINT DE VUE UTILISATEUR DE LA CONTROVERSE SUR TWITTER À L'AIDE DE RÉSEAUX NEURONAUX GRAPHIQUES

Sommaire

5.1 Introduction

Dans ce chapitre, nous présentons nos travaux effectués sur la quantification de la controverse sur Twitter. Nous basons nos mesures de la controverse sur les comportements de l'utilisateur, capturés via leurs productions textuelles et/ou leurs actions (retweets). Nous présentons une modélisation théorique d'un score de quantification. Nous proposons ensuite une méthode afin d'estimer ce score empiriquement, en moyennant la probabilité d'un utilisateur d'appartenir à un sujet controversé. Comme dans le chapitre 4, notre méthode se base sur les GNN comme modèles d'apprentissage, et utilise les informations textuelles des tweets et structurelles du graphe des utilisateurs de retweets. Cependant, contrairement au chapitre 4, la représentation textuelle des utilisateurs est affinée grâce à l'apprentissage de nouvelles représentations au sein du modèle. Ces travaux sont encore en cours de développement au moment de la rédaction de ce manuscrit, mais les premiers résultats présentés sont prometteurs.

Le reste du chapitre est organisé comme suit. La section 5.2 introduit le contexte de nos travaux, ainsi que notre approche et les contributions réalisées. La section 5.3 présente tout d'abord la modélisation théorique de notre score de quantification de la controverse sur Twitter et présente ensuite notre méthode afin d'estimer empiriquement ce score. La section 5.4 regroupe les différents estimateurs traités durant nos expériences, ainsi que le protocole d'évaluation mis en place afin de les comparer. Les travaux n'étant pas finis, la section 5.5 regroupe les premiers résultats obtenus. Enfin, la section 5.6 recense les travaux encore à effectuer, les perspectives envisagées, et conclut le chapitre.

5.2 Contexte

De la même manière que dans le chapitre 3, nous traitons dans ce nouveau chapitre de la controverse du point de vue de la polarisation des utilisateurs autour de communautés aux points de vue et opinions différentes dans les réseaux sociaux. Nous nous focalisons sur le réseau social Twitter et les graphes de retweets entre utilisateurs. La quantification de la controverse est une tâche difficile, dû notamment à la subjectivité du concept. La quantification de la controverse selon la polarisation des communautés a été beaucoup étudiée dans la littérature [Gar+18a ; Mor+15]. La plupart des approches se concentrent seulement sur cette polarité, en ne prenant pas en compte les informations contenues dans les textes publiés par les utilisateurs.

Contributions. Dans ce chapitre, nous nous intéressons donc à la quantification de la controverse sur Twitter. Pour cela, nous proposons une approche basée sur les GNN et la prédiction de la participation d'un utilisateur à un sujet controversé, résultant des trois contributions suivantes :

- **Quantification de la controverse**. Nous proposons une modélisation théorique afin de quantifier la controverse en fonction des utilisateurs concernés par le sujet. Ce score se base sur l'espérance de la probabilité d'un utilisateur de participer à un sujet controversé.
- **Estimation empirique du score de quantification.** Nous proposons ensuite une méthode afin d'estimer ce score empiriquement. Pour cela, nous utilisons les

probabilités issues de l'estimateur qui évalue si les utilisateurs participent à un sujet controversé. Nous démontrons que minimiser la fonction de perte de cet estimateur revient à optimiser notre score. Afin d'estimer cette probabilité pour chaque utilisateur, nous proposons, comme dans le chapitre 4, un modèle basé sur les GNN, exploitant à la fois les informations textuelles et structurelles du graphe.

- **Résultats et analyse**. Nous établissons un protocole d'évaluation pour comparer nos différents modèles. Nous présentons les premiers résultats obtenus, en comparant notre modèle avec des variantes de celui-ci utilisant uniquement le texte ou la structure. Puis, en comparant les scores de quantification avec ceux de la littérature sur des sujets réels provenant de Twitter, nous montrons que notre approche réalise de meilleures performances. La suite des expérimentations envisagées est aussi présentée, ainsi que des améliorations au niveau de l'estimation du modèle, à l'aide de sa calibration.

5.3 Méthodologie

5.3.1 Quantification de la controverse

Dans cette section, nous proposons un score théorique de quantification de la controverse à partir des sous-graphes utilisateurs.

Soit G un graphe aléatoire et L un label aléatoire. On note g et l leur réalisation. On indicera parfois le graphe et les labels comme dans le chapitre 4 (g_i et l_i) lorsque nécessaire. Soit $\mathbb{P}$ la loi jointe de G, L qui se décompose en μ, la marginale de G, et η la loi de L conditionnellement à l'observation d'un graphe [DGL13]. $\mathbb{P}$ représente le processus générateur tel que $G, L \sim \mathbb{P}$. Ce processus générateur est inconnu. Soit η la probabilité du label conditionnellement à l'observation d'un graphe :

$$\eta(g) = \mathbb{P}(L = 1 | G = g) \tag{5.1}$$

L'équation 5.1 correspond à la probabilité réelle qu'un graphe soit controversé ($L = 1$) conditionnellement à l'observation du graphe ($G = g$). On appelle parfois η la *posterior* en opposition au *prior* qui représente la fréquence des labels.

Notre score de quantification de la controverse est lié à la manière dont les caractéristiques de cette dernière apparaissent dans le graphe utilisateur. Ainsi, si toute partie du graphe implique nécessairement la controverse, alors le score de quantification devrait être élevé. À l'inverse, si seulement certaines parties du graphe permettent de déterminer que le sujet est controversé, alors, le score devrait être faible. Enfin, si le graphe n'est pas associé à un sujet controversé, le score devrait être proche de 0.

Afin de définir ce score, notons $g_u^{(k)} \subseteq g$ le sous-graphe de g centré sur l'utilisateur u, allant jusqu'à k niveaux de voisinage.

La probabilité conditionnelle (selon la vraie loi) que le sous-graphe $g_u^{(k)}$ soit associé à un sujet controversé est donnée par l'équation 5.2.

$$\eta\left(g_u^{(k)}\right) = \mathbb{P}\left(L = 1 | G = g_u^{(k)}\right) \tag{5.2}$$

$\eta\left(g_u^{(0)}\right)$ représente la probabilité que le contenu publié par l'utilisateur u soit associé à un sujet controversé indépendamment de toute interaction. À l'opposé, $\eta\left(g_u^{(\infty)}\right)$ représente la probabilité que le sujet soit controversé en analysant la totalité du graphe. On s'attend évidemment à ce que cette dernière soit proche de 1, même si parfois l'absence d'éléments de contexte peut limiter la certitude de prédiction. Notons que cette dernière reste une quantité qui dépend de η qui est notre *posterior* inconnue.

Nous définissons maintenant le score de quantification CQS (pour "Controversy Quantification Score" en anglais) :

$$CQS(g, k) = \mathbb{E}_{u\sim\mathbb{U}(g)}\left[\eta\left(g_u^{(k)}\right)\right] \tag{5.3}$$

Ce score correspond à l'espérance qu'un utilisateur choisi uniformément dans le graphe soit associé à un sujet controversé, sachant que l'on observe uniquement k niveaux de voisinage.

La seule quantité inconnue de l'équation 5.3 est la loi de probabilité conditionnelle, ou *posterior*, η qu'il convient d'estimer.

5.3.2 Fonctions de perte consistantes avec CQS

Dans cette section, nous montrons que les fonctions de pertes (ou "*loss*" en anglais) souvent utilisées en *deep learning*, telle que la *cross entropy*, sont consistantes avec l'estimation de notre score CQS. Dit autrement, la minimisation d'une fonction de perte consistante minimise l'erreur d'estimation de CQS.

Soit $\ell : \{0, 1\} \times [0, 1] \to \mathbb{R}^+$ une fonction de perte binaire. Le premier argument est le label (1 ou 0 en fonction de l'état controversé ou non) et le second est une estimation de la probabilité du label. Soit η la probabilité réelle (selon la vraie loi de probabilité) et $\hat{\eta}$ son estimation. Le risque associé à ℓ est donné par :

$$\mathcal{L}_\ell(\hat{\eta}, \eta) = \eta\ell\left(\hat{\eta}\right) + (1 - \eta)\ell\left(1 - \hat{\eta}\right) \tag{5.4}$$

Si $\ell(s) = -\log(s)$, alors on obtient la *binary cross entropy*.

Définition 5.1 ((*Strictly*) *Proper loss* [Lor20])**.** Une loss $\ell : \{0, 1\} \times [0, 1] \to \mathbb{R}^+$ est *proper* si son infimum (plus grand minorant) est atteint par η :

$$\mathcal{L}_\ell(\eta, \eta) = \inf_{s\in[0,1]} \mathcal{L}_\ell(s, \eta)$$

Et *strictement proper* si η est l'unique minimiseur.

Définition 5.2 (*μ-strongly proper loss* [Lor20])**.** Une loss $\ell : \{0,1\} \times [0,1] \to \mathbb{R}^+$ est *μ-strongly proper* si :

$$\mathcal{L}_\ell(\hat{\eta}, \eta) - \mathcal{L}_\ell(\eta, \eta) \geq \frac{\mu}{2}|\hat{\eta} - \eta|_1^2$$

Le regret est l'écart entre notre risque et sa valeur optimale :

$$\text{Reg}_\ell(\hat{\eta}; x) = \mathcal{L}_\ell(\hat{\eta}(x), \eta(x)) - \mathcal{L}_\ell(\eta(x), \eta(x))$$

La proposition suivante montre qu'une fonction de perte *strongly proper* est consistante avec notre score si les graphes sont générés selon une procédure spécifique :

$$\begin{aligned} G &\sim \mu \\ u &\sim \mathbb{U}(G) \end{aligned} \tag{5.5}$$

d'où on construit le graphe $g_u^{(k)}$. μ est la marginale des graphes.

Proposition 5.1. *Toute fonction de perte ℓ strongly proper est consistante avec CQS si elle est minimisée pour une génération des graphes selon la procédure ci-dessus.*

$$\mathbb{E}_G\left[\mathbb{E}_{u\sim\mathbb{U}(G)}\left[Reg_\ell(\hat{\eta}; g_u^{(k)})\right]\right] \to 0 \Rightarrow \mathbb{E}_G\left[|\widehat{CQS}(G,k) - CQS(G,k)|\right] \to 0$$

Démonstration.

$$\begin{aligned} &\mathbb{E}_G\left[|\widehat{CQS}(G,k) - CQS(G,k)|\right] \\ &\quad= \mathbb{E}_G\left[|\mathbb{E}_{u\sim\mathbb{U}(G)}\left[\hat{\eta}(g_u^{(k)}) - \eta(g_u^{(k)})\right]|\right] \\ &\quad\leq \mathbb{E}_G\left[\mathbb{E}_{u\sim\mathbb{U}(G)}\left[|\hat{\eta}(g_u^{(k)}) - \eta(g_u^{(k)})|\right]\right] \text{ (inégalité de Jenssen)} \\ &\quad\leq \sqrt{\frac{2}{\mu}\mathbb{E}_G\left[\mathbb{E}_{u\sim\mathbb{U}(G)}\left[\text{Reg}\ell(\hat{\eta}; g_u^{(k)})\right]\right]} \text{ (car } \textit{strongly proper} \text{ [Lor20])} \end{aligned} \tag{5.6}$$

Ainsi, si ce dernier terme converge vers 0, alors l'erreur d'estimation de notre score aussi.

□

On peut généraliser le théorème précédant en prenant un k aléatoire selon une loi de probabilité. Si le regret tend vers 0 pour un k aléatoire, alors l'erreur d'estimation aussi. Ainsi, on pose alors, avec μ_1 la loi de probabilité marginale de k et μ_2 la loi de probabilité marginale des graphes :

$$k \sim \mu_1$$

$$\begin{aligned} g_i &\sim \mu_2 \\ u &\sim U(g^{(k)}) \end{aligned}$$

Nous présentons ensuite dans la section 5.3.3 une méthode pour estimer ce score de quantification de manière empirique.

5.3.3 Estimation empirique des probabilités conditionnelles

Après avoir défini le score $CQS(g, k)$ de manière théorique dans l'équation 5.3, nous établissons une méthode afin d'estimer ce score de manière empirique. Pour cela, on calcule les probabilités conditionnelles $\eta(g_u^{(k)})$ à partir des sous-graphes à k niveaux de chaque utilisateur u. La figure 5.1 représente les différentes étapes du processus de calcul de cette probabilité, et de l'estimation de notre score de quantification de la controverse.

Nous présentons maintenant notre modèle basé sur les GNN afin d'estimer η. Tout d'abord, le graphe utilisateur de retweets, présenté dans la section 5.3.3.1, est construit à partir des tweets récupérés en rapport avec les sujets. Le graphe est ensuite transmis à notre modèle, afin de prédire la participation de l'utilisateur à un sujet controversé. La section 5.3.3.2 présente ce modèle, qui combine à la fois les propriétés structurelles du graphe de retweets et les informations textuelles publiées par les utilisateurs dans leurs tweets originaux. Contrairement au chapitre 4, la représentation textuelle des utilisateurs est affinée grâce à l'apprentissage de nouvelles représentations au sein du modèle.

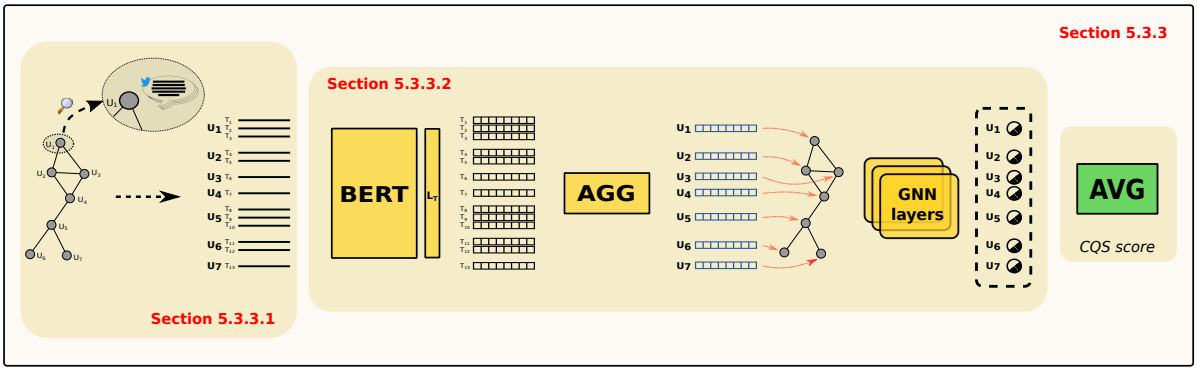

FIGURE 5.1 – *Vue d'ensemble des différentes étapes de notre approche afin de quantifier la controverse d'un point de vue utilisateur. Contrairement au chapitre 5, la représentation textuelle des utilisateurs est affinée grâce à l'apprentissage de nouvelles représentations dans le modèle. La sortie de la dernière couche de GNN est une estimation de la probabilité d'un utilisateur de participer à un sujet controversé.*

5.3.3.1 Construction du graphe utilisateur

Comme expliqué dans les chapitres précédents, nous supposons que le retweet représente l'approbation, et permet de construire des communautés avec des opinions partagées. Comme dans la section 4.5.1, nous construisons un graphe utilisateurs de retweets $g_u^{(k)}$ représentant le sujet, centré sur l'utilisateur u, allant jusqu'à k niveaux de voisinage. Deux utilisateurs u_i et u_j sont reliés par un lien si l'un a retweeté l'autre au moins une fois. Le graphe est non dirigé. Les graphes construits sont de la même forme que ceux présentés dans la figure 2.2. Chaque utilisateur est représenté par ses propres tweets originaux dans le sujet concerné. Une représentation tokenisée est générée à partir du "tokenizer" BERT (cf. section 4.3.3) pour chacun des tweets.

Plus formellement, un sujet t est représenté par un graphe $G = (\mathcal{U}, \mathcal{E}, X)$ où $\mathcal{U} = \{u_1, u_2, ..., u_n\}$ désigne les nœuds utilisateurs et $\mathcal{E} = \{(u_i, u_j)\}_{1 \leq i,j \leq n}$ représente les arêtes du graphe. Chaque nœud correspond à un utilisateur unique. Une arête entre deux nœuds existe s'il y a interaction entre ces deux utilisateurs.

5.3.3.2 Prédiction de la participation des utilisateurs à un sujet controversé

Chaque tweet est représenté par un vecteur à l'aide du modèle de langage BERT. La sortie de la dernière couche ajoutée à BERT sert de vecteur de représentation pour chaque tweet. Pour représenter chaque utilisateur, une fonction d'agrégation est utilisée pour regrouper les représentations des tweets, indiquée dans la figure 5.1 par le bloc AGG. Contrairement au chapitre 4, la représentation des tweets est affinée lors de l'entrainement du modèle. Ainsi, la représentation textuelle des utilisateurs après la fonction d'agrégation est affinée. Les utilisateurs n'ayant pas posté de tweets se voient attribuer un tweet vide par défaut, avec sa représentation vectorielle correspondante. Tous les vecteurs de représentation des utilisateurs sont ensuite regroupés dans la matrice $X \in \mathbb{R}^{n \times d}$, avec n le nombre de nœuds et d la dimension des vecteurs de représentation.

À partir de la représentation des tweets de chaque utilisateur, le modèle va maintenant apprendre une nouvelle représentation intégrant la représentation structurelle du graphe. Pour apprendre cette représentation, un GNN à couches multiples est inclus dans le modèle, comme dans le chapitre 4. Cette étape est représentée par les blocs "GNN layers" dans la figure 5.1. Nous présentons les deux types de convolutions étudiées.

- **Convolution inductive pour la représentation des nœuds**. La méthode de convolution $GraphSAGE$ [HYL17] utilise l'échantillonnage et l'agrégation du voisinage pour générer de nouvelles représentations de nœuds à partir des voisins locaux. Comparé à la théorie spatiale classique des couches convolutives [Xu+19], évoquée dans le chapitre 2, $GraphSAGE$ utilise l'échantillonnage de nœuds afin de maintenir la taille du voisinage fixe. Le fonctionnement de cette approche est expliqué en détails dans la section 2.3.2.2.
- **Convolution sur la base du mécanisme d'attention**. La méthode de convolution GAT [Vel+18] apprend les nouvelles représentations des nœuds du graphe en tirant parti des mécanismes d'attention. Le mécanisme d'auto-attention est introduit pour attribuer des poids d'attention différents au voisinage des utilisateurs dans le graphe de retweets. En apprenant l'importance des voisins de chaque utilisateur, GAT peut se concentrer sur les utilisateurs les plus pertinents lors de l'apprentissage de la nouvelle représentation de l'utilisateur cible. Le fonctionnement de cette approche est expliqué en détails dans la section 2.3.2.2.

Ces deux types de convolution ont été retenus pour leur capacité inductive, ainsi que l'utilisation d'un nombre restreint de nœuds lors du calcul des nouvelles représentations. Les coûts de calcul et de mémoire du modèle sont alors réduits, le graphe entier n'étant pas nécessaire lors du calcul de la nouvelle représentation d'un nœud.

Seul le nœud et son voisinage sont nécessaires. Dans notre modèle, la dernière couche convolutive sera utilisée pour classer le nœud utilisateur comme appartenant à un sujet controversé ou non, en utilisant la fonction $SOFTMAX$ comme fonction d'activation. Notons que pour la première couche $l = 0$, les caractéristiques d'entrée des utilisateurs correspondent à la représentation textuelle X des utilisateurs. Il est important de noter que contrairement au chapitre 4, la représentation textuelle des utilisateurs est ainsi affinée pendant l'apprentissage de nouvelles représentations par le modèle.

Notre estimateur $\hat{\eta}$ étant défini, nous pouvons à présent estimer notre score de quantification en prédisant pour chaque utilisateur u sa probabilité de participer à un sujet controversé, puis en estimant empiriquement notre score de quantification en moyennant toutes ces probabilités obtenues.

5.4 Évaluations des expérimentations

Dans cette section, nous présentons les expérimentations réalisées afin d'évaluer notre approche. Dans un premier temps, nous présentons dans la section 5.4.1 le protocole d'évaluation mis en place afin d'évaluer la qualité des différents modèles présentés à prédire la probabilité d'un utilisateur de participer à un sujet controversé. Ce protocole est utilisé afin de comparer les performances de nos modèles dans la section 5.5.1. Dans un second temps, nous évaluons notre approche en faisant varier différentes caractéristiques et paramètres, que nous présentons dans la section 5.4.2 et en la comparant à d'autres approches de la littérature dans la section 5.5.2. Comme dans les chapitres 3 et 4, nous analysons aussi l'apport de la combinaison des caractéristiques structurelles et textuelles.

Le jeu de données utilisé comprend quinze sujets controversés et quinze sujets non controversés provenant de Twitter, présentés dans la section 2.4.3. Chaque sujet comprend plusieurs milliers de tweets et de retweets. Des statistiques détaillées sur chacun de ces sujets sont présentées dans la table 2.4.

5.4.1 Protocole d'évaluation

Nous présentons dans cette section le protocole d'évaluation afin de juger de la qualité de nos différents modèles. Pour entraîner et tester notre modèle, le jeu de données est divisé en deux ensembles équilibrés, l'un pour l'entraînement et l'autre pour le test. Le jeu d'entrainement $\mathcal{G}_{\text{train}}$ contient 20 sujets (10 de chaque label), et le jeu de test $\mathcal{G}_{\text{test}}$ 10 sujets (5 de chaque label). Afin d'éviter de biaiser notre analyse avec un sur-apprentissage du modèle, nous veillons à ce que les sujets controversés séparés par période de temps (BOLSONARO{XXX}, MENCIONES{XXX}, KAVANAUGH{XXX}) fassent partie du même jeu (d'entraînement ou de test).

Comme présenté dans la partie théorique, afin de tester et comparer nos approches, nous devons définir une métrique. Pour cela, nous allons créer plusieurs sous-jeux de tests selon la valeur de k. Chaque sous-jeu de test comprendra des sous-graphes centrés sur chaque utilisateur avec leurs voisins à k niveau, pour chacun des graphes compris

dans le jeu de test. Les sous-graphes pour $k = 0$ correspondent à un nœud (l'utilisateur) décrit par ses messages. Or, la plupart des utilisateurs ne publient pas de tweets, mais des retweets, d'où la nécessité de prendre aussi en compte le voisinage. Au contraire, les sous-graphes pour $k = +\infty$ correspondent pour chaque utilisateur au graphe entier. En théorie, on définit un niveau k, à partir duquel un utilisateur participe à un sujet controversé. Cette valeur de k est difficile à choisir sans connaissance sociologique et philosophique de la controverse. Ce k peut également varier d'un sujet à l'autre. Nous analysons donc les performances de nos modèles pour différentes valeurs de k.

Afin d'évaluer les performances et comparer nos modèles, nous définissons un jeu de test en créant des sous-graphes $g_u^{(k)}$ à partir des graphes compris dans $\mathcal{G}_{\text{test}}$. pour chaque utilisateur u de chaque sujet en test, avec des valeurs de k aléatoires. k suit une loi de Poisson avec un $\lambda = 2$, afin d'avoir un niveau de voisinage moyen de 2, correspondant à un bon équilibre entre petits et grands graphes. Nous analyserons néanmoins l'évolution de nos performances en créant des sous-jeux de test, pour notre meilleur modèle, selon différentes valeurs de k. La fonction de *cross-entropy*, rapportée dans l'équation 5.7, est utilisée comme métrique afin de tester et comparer les performances de nos modèles.

$$\begin{aligned}\mathcal{L}(\hat{\eta}) &= \mathbb{E}_{G,L}\left[\mathbb{E}_{u\sim\mathbb{U}(G)}\left[-L\log\left(\hat{\eta}\left(g_u^{(k)}\right)\right) - (1-L)\log\left(1-\hat{\eta}\left(g_u^{(k)}\right)\right)\right]\right] \\ &\approx -\frac{1}{|\mathcal{G}_{\text{test}}|}\sum_{i=1}^{|\mathcal{G}_{\text{test}}|}\frac{1}{|g_i|}\sum_{u\in g_i} l_i\log\left(\hat{\eta}\left(g_{i,u}^{(k)}\right)\right) + (1-l_i)\log\left(1-\hat{\eta}\left(g_{i,u}^{(k)}\right)\right)\end{aligned} \tag{5.7}$$

$|\mathcal{G}_{\text{test}}|$ représente le nombre de graphes dans notre jeu de test, et $|g_i|$ le nombre de nœuds dans le graphe g_i. La variable l_i correspond au label du graphe g_i depuis lequel le sous-graphe $g_{i,u}^{(k)}$ a été construit. L'objectif de nos modèles est de minimiser cette fonction, afin de maximiser le score quantification.

5.4.2 Présentation des différents modèles

Nous présentons ici les différents modèles et types d'entrainement testés. Tout d'abord, nous comparerons deux types d'entrainement :

- Dans le premier cas, pour chacun des sujets compris dans le jeu d'entrainement, le graphe complet est utilisé à chaque époque.
- Dans le deuxième cas, pour limiter le problème de manque de données (uniquement 20 sujets dans le jeu d'entraînement), nous effectuons une étape d'augmentation des données, en échantillonnant des sous-graphes centrés sur les utilisateurs, à différents niveaux k de voisinage. Cela permet au modèle de voir des graphes de différentes tailles.

Pour ces deux types d'entraînement, nous testons différentes combinaisons de paramètres. Nous utilisons un modèle basé sur BERT, avec une couche supplémentaire de dimension 768, pour obtenir les représentations des tweets. Pour réduire les coûts de

calcul et de temps, seuls les poids de la couche supplémentaire sont mis à jour pendant la phase d'apprentissage. $MEAN$ est utilisé comme agrégateur pour représenter les utilisateurs à partir de leurs tweets. Enfin, concernant le modèle GNN, nous testons les deux modèles présentés dans la section 5.3.3.2 (GAT et $GraphSAGE$) avec 1, 2 et 3 couches de convolution de dimensions 192, afin de comparer la représentation finale locale et globale des utilisateurs. Les modèles sont entrainés avec un taux d'apprentissage de 1×10^{-3} avec un poids décroissant ("weigh decay" en anglais) de 0,05, et une taille de lot de 64. Les modèles sont formés pendant un maximum de 300 époques, avec un arrêt précoce si la fonction de perte ne diminue pas après 100 époques. Les modèles sont optimisés à l'aide de l'optimiseur $ADAM$ appliqué sur la fonction de perte *cross-entropy*. Au cours de la phase d'apprentissage, nous échantillonnons et prenons aléatoirement un nombre fixe de nœuds, correspondant à la taille du plus petit des graphes d'entrainements, pour chacun des sujets afin d'assurer un apprentissage équilibré et d'élargir le champ des utilisateurs observés. Les modèles GNN utilisés dans cette étude étant inductif, un batch d'entrainement correspond à un seul graphe comprenant tous les nœuds utilisateurs échantillonnés des graphes d'entrainement.

Pour comparer les performances de nos différents modèles utilisant à la fois des informations structurelles et textuelles, nous avons défini deux modèles de référence. "$\text{GRAPH}_{\text{DEGREE}}$" se base sur les GNN, mais utilise uniquement des informations structurelles. Nous utilisons la même méthodologie que notre approche, mais à la place des caractéristiques textuelles en entrée, nous utilisons le degré du nœud comme caractéristique de l'utilisateur. "$\text{TEXT}_{\text{BERT}}$" se base sur un modèle BERT, utilisant uniquement les informations textuelles de l'utilisateur pour prédire la participation à un sujet controversé. Le modèle BERT est affiné en utilisant les tweets originaux, étiquetés appartenant à un sujet controversé ou non. Chaque utilisateur dispose d'un ensemble de caractéristiques correspondant à tous ses tweets et retweets. Nous considérons ici de la même manière un tweet et un retweet. La prédiction finale de l'utilisateur correspond à la probabilité moyenne de chacun de ses tweets (et retweets) prédite par le modèle BERT. Ce modèle est entraîné en utilisant un taux d'apprentissage de 2×10^{-5} et une taille de lot de 64. Le modèle est entraîné pendant 30 époques au maximum, s'arrêtant si la fonction de perte ne diminue pas après 3 époques, et optimisé à l'aide de l'algorithme $ADAM$ sur la fonction de perte *cross-entropy*.

5.5 Résultats pour les sous-graphes au k maximum

Au moment d'écrire ce mémoire, les expérimentations présentées dans la section 5.4 sont en cours. Dans cette section, nous présentons les premières expérimentations effectuées et les premiers résultats obtenus. Nous utilisons ici les sous-graphes pour $k = +\infty$, représentant chaque utilisateur avec son maximum de voisins possible pour calculer sa probabilité de participer à un sujet controversé.

5.5.1 Comparaison des différents modèles

Nous traitons pour l'instant seulement le cas de l'entraînement en utilisant les graphes complets. Ces expérimentations ont été menées sur les machines *Jean-Zay* de

TABLE 5.1 – *Comparaison des performances de nos modèles pour la prédiction de la participation d'un utilisateur à un sujet controversé avec les modèles de référence. Les performances sont évaluées avec la cross-entropy, sur le sous-jeu de test correspondant aux sous-graphes centrés, pour $k = +\infty$ niveaux de voisinages.*

$k = +\infty$	**Graphes complets**
	cross-entropy
$\text{GRAPH}_{\text{DEGREE}}$	0,794
$\text{TEXT}_{\text{BERT}}$	0,579
GRAPHSAGE_MEAN_1	1,086
GRAPHSAGE_MEAN_2	1,1
GRAPHSAGE_MEAN_3	0,692
GAT_MEAN_1	0,359
GAT_MEAN_2	**0,323**
GAT_MEAN_3	1.234

l'IDRIS.

La table 5.1 regroupe les performances (évaluées par la *cross-entropy* sur le jeu de test $k = +\infty$) obtenues. Les expériences montrent clairement que l'utilisation des réseaux d'attention (GAT) sur les informations structurelles et textuelles des utilisateurs, combinées à l'agrégateur $MEAN$, pour la prédiction de la participation d'un utilisateur à un sujet controversé, permet d'obtenir de meilleures performances. Nos expériences montrent que l'utilisation d'un voisinage trop global (3 couches de convolution) lors de l'apprentissage des nouvelles représentations utilisateurs conduit à un sur-apprentissage du modèle, avec une *cross-entropy* nettement plus élevée pour le modèle GAT_MEAN_3. Notre modèle utilisant deux couches de convolution ("GAT_MEAN_2") obtient les meilleures performances globales, avec une cross-entropy de 0,323, comparé au modèle utilisant uniquement les informations de voisinage local avec une couche de convolution (*cross-entropy* à 0,359). Les modèles basés sur $GraphSAGE$ ne parviennent pas à prédire correctement la participation des utilisateurs à un sujet controversé, quel que soit le type d'agrégateur et le nombre de couches. L'utilisation des poids d'attention sur le voisinage permet d'obtenir des représentations plus précises des utilisateurs et un meilleur apprentissage du modèle. Nos expériences montrent aussi que GAT_MEAN_2, combinant informations structurelles et textuelles, obtient de meilleurs résultats que ceux du modèle utilisant uniquement les informations textuelles ($\text{TEXT}_{\text{BERT}}$ a une *cross-entropy* de 0,579). Elle montre également que l'utilisation des seules caractéristiques structurelles n'obtient pas de meilleures performances ($\text{GRAPH}_{\text{DEGREE}}$ a une *cross-entropy* de 0,794). Comme dans le chapitre précédent, cela conforte notre intuition que les caractéristiques textuelles et structurelles contiennent différentes informations complémentaires sur la controverse.

Malgré des premiers résultats assez intéressants, le reste des expérimentations, ainsi que l'analyse de l'évolution des performances pour différentes valeurs de k en jeu de test, restent donc à effectuer.

TABLE 5.2 – *Comparaison de l'aire sous la courbe roc_auc des scores calculés à partir de modèles utilisant différentes caractéristiques. Les roc_auc sont calculées à partir des scores estimés sur les sujets appartenant au sous-jeu de test pour $k = +\infty$.*

Baseline		**CQS**		
rwc_score	$dipole\_score$	$\text{GRAPH}_{\text{DEGREE}}$	$\text{TEXT}_{\text{BERT}}$	GAT_MEAN_2
0,76	0,8	0,84	0,92	**1,0**

5.5.2 Comparaison des scores de quantification avec la littérature

En se basant sur ces premiers résultats, on peut donc calculer l'estimation de nos scores de controverse sur notre jeu de test, et les comparer aux méthodes de la littérature. Comme indiqué dans la section 5.3.1, le modèle estimant le mieux la probabilité de participation d'un utilisateur à la controverse dans les sous-graphes k, implique un meilleur score de quantification.

En reprenant la formule de notre score de quantification, on calcule $g_u^{(+\infty)}$ pour chaque utilisateur du graphe, avec $\hat{\eta}$ représenté par notre modèle obtenant les meilleures performances, GAT_MEAN_2.

Nous regardons ensuite si notre score de quantification permet une bonne séparation des sujets controversés et non controversés. Pour cela, nous évaluons, à l'aide du score roc_auc, la capacité des scores à discerner les classes en mesurant l'aire sous la courbe ROC. Nous analysons visuellement aussi l'estimation par noyau ("kernel density" en anglais) de la densité des distributions selon le label des sujets pour chaque score. Nous comparons notre score avec les deux scores de polarisation de la controverse présentés par GARIMELLA et al. [Gar+18a] : le rwc_score basé sur l'échantillonnage de marches aléatoires, et le $dipole\_score$ basé sur la répartition et l'alignement des charges électriques (nœuds) dans une molécule (graphe). Nous comparons aussi notre score avec ce même score calculé à partir des prédictions des modèles utilisant uniquement les caractéristiques structurelles ou textuelles : $\text{GRAPH}_{\text{DEGREE}}$ et $\text{TEXT}_{\text{BERT}}$. Notre comparaison porte sur les 10 sujets du jeu de test $\mathcal{G}_{\text{test}}$.

La table 5.2 regroupe les roc_auc de tous les scores. Notre score CQS calculé à partir des probabilités prédites par le modèle GAT_MEAN_2 obtient de meilleurs résultats que les scores basés seulement sur la structure ou sur les tweets, qu'ils proviennent de nos modèles ou des scores de la littérature sur les 10 sujets du jeu de test. Notre approche apporte donc des informations supplémentaires dans la manière de discerner les sujets.

Dans la figure 5.2, on remarque au niveau des densités de distribution que notre score entraine moins de chevauchements entre la courbe des sujets controversés et celle des non controversés, en comparaison des autres scores. De plus, les courbes de distribution selon le label de notre score CQS basé sur les prédictions du modèle GAT_MEAN_2, sont bien reparties autour du score moyen[1] $(x = 0,5)$, comparé aux

1. L'intervalle de notre score de quantification se trouve entre de 0 à 1, 0 indiquant que le sujet n'est pas du tout controversé, et 1 très controversé.

scores CQS provenant des modèles $\text{GRAPH}_{\text{DEGREE}}$ et $\text{TEXT}_{\text{BERT}}$.

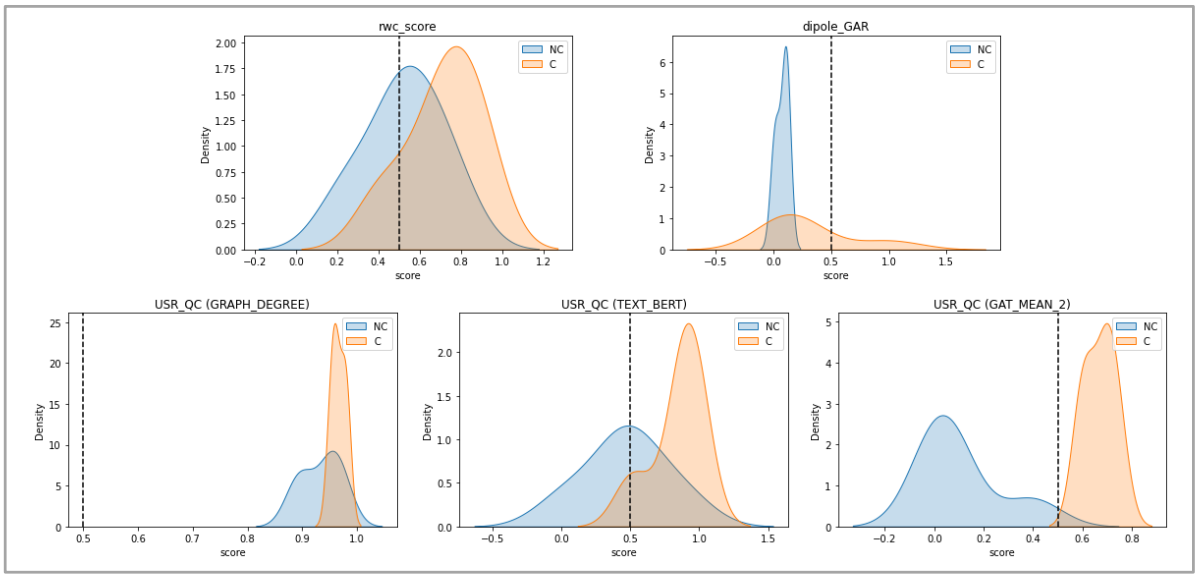

Figure 5.2 – *Diagramme des distributions de l'estimation par noyau selon le label des sujets, pour chacun des scores. Les deux diagrammes en haut correspondent aux scores de notre base de référence. Les trois diagrammes en dessous correspondent à des scores basés sur différents modèles, dont celle de notre approche. La courbe bleue correspond à la distribution des sujets non controversés* (NC)*, et la courbe orange correspond à celle des sujets controversés* (C)*.*

Notre approche discerne bien les sujets controversés des non controversés, mais ne garantit en rien la qualité de la quantification. Cette dernière reste très compliquée à juger, du fait de la subjectivité d'un tel concept et du besoin d'annotateurs humains garantissant l'exactitude d'un score pour chaque sujet. Afin d'étudier et de vérifier la cohérence de notre score, l'analyse reste aussi encore à compléter, au gré de l'avancement des expérimentations présentes dans la section 5.4.

5.6 Conclusions et Perspectives

Dans ce chapitre, nous avons présenté une méthode afin de quantifier la controverse d'un sujet sur Twitter. Nous avons présenté une modélisation théorique de notre score de quantification. Nous démontrons que minimiser la fonction de perte de notre estimateur revient à optimiser le score de quantification, et présentons ensuite notre méthode afin d'estimer ce score empiriquement. Cette probabilité est calculée à partir d'un modèle basé sur les GNN, permettant ainsi d'utiliser à la fois les caractéristiques structurelles du graphe utilisateur et textuelles des tweets. Nous avons aussi présenté un protocole d'évaluation, basé sur des sous-jeux de test comprenant des sous-graphes, centrés sur les utilisateurs des sujets en test, avec des niveaux de voisinage différents. Les premières expérimentations sont concluantes, même si ces travaux sont encore en cours. Sur le jeu de test comprenant des graphes entiers $(k = +\infty)$, notre modèle obtient de meilleures performances lorsque l'on utilise les deux types de caractéristiques

structurelles et textuelles, plutôt que séparément, confortant notre intuition que les interactions entre utilisateurs et le contenu de leurs tweets contiennent des informations complémentaires sur la controverse. Nous devons encore évaluer ce score sur des jeux de test à différents niveaux de voisinage k. Notre approche présente l'avantage de ne pas être dépendante du nombre d'utilisateurs utilisés dans le calcul de quantification. En plus du gain de temps de calcul et de mémoire, des perspectives intéressantes existent, comme l'étude de l'impact d'un groupe d'utilisateurs sur la controverse.

Pour finir, nous présentons une perspective importante portant sur la calibration de notre modèle de quantification. Actuellement, afin d'estimer la probabilité η d'un utilisateur de participer à un sujet controversé, nous utilisons l'estimation faite par notre modèle sur sa dernière couche de prédiction, à l'aide de la fonction logistique $SOFTMAX$. Cependant, les réseaux de neurones modernes sont reconnus pour avoir une faible capacité à estimer la probabilité de confiance dans leur prédiction. Ils sont souvent trop confiants, avec des scores $SOFTMAX$ proches de 100%. Afin de résoudre ce problème, des solutions de calibration du modèle sont mises en place [Guo+17; GT22]. La calibration d'un réseau de neurones est un processus garantissant que les prédictions du modèle correspondent aux probabilités réelles des événements. Elle implique l'ajustement des scores de confiance prédits par le modèle pour qu'ils reflètent avec précision la fiabilité des prédictions. Différentes possibilités peuvent être mises en place dans notre cas :

- Ajustement du $SOFTMAX$: la première solution serait d'ajuster le $SOFTMAX$, en utilisant la méthode d'étalonnage "Temperature Scaling" [Kul+19]. Elle ajuste la "température" de la distribution de probabilités pour rendre les prédictions du modèle plus conformes aux probabilités réelles, afin d'améliorer la fiabilité des prédictions.
- Utilisation d'ensemble de modèles : cette solution combine les prédictions de plusieurs modèles pour produire une prédiction étalonnée plus précise et fiable. Le principal problème reste le coût de calcul, proportionnel au nombre de modèles utilisés.
- Ajout de la fonction "monte-carlo dropout" [GG16] : cette solution utilise une fonction de "dropout" lors de la phase d'entrainement. Pendant la phase d'inférence, elle désactive aléatoirement des neurones dans le modèle. En moyennant les résultats de ces prédictions, il est possible d'estimer l'incertitude associée à chaque prédiction, permettant de mieux quantifier la fiabilité des prédictions du modèle.

Ensuite, pour comparer ces méthodes de calibration sur nos jeux de données, nous pouvons calculer le "Brier score" [Ruf10], fonction mesurant l'exactitude des prédictions probabilistes. La calibration de notre modèle nous permettra d'affiner notre score de quantification, en évitant notamment une surestimation de ces scores.

Ce chapitre présente les derniers travaux en cours durant cette thèse. Le chapitre suivant conclut les travaux effectués lors de cette thèse, et présente les perspectives à venir.

CHAPITRE 6

CONCLUSIONS ET PERSPECTIVES

Sommaire

Ce chapitre conclut ce manuscrit de thèse. Il présente un bilan général de nos contributions scientifiques dans le domaine de la controverse. Il dresse ensuite des perspectives de recherche qui pourraient être menées en vue d'améliorer les tâches d'explication, de détection et de quantification de la controverse, telles que définies dans ce manuscrit.

6.1 Conclusions

Dans l'ensemble des chapitres, nous avons exploré différentes approches pour expliquer, détecter et quantifier la controverse. Nous avons principalement utilisé les GNN pour capturer les caractéristiques structurelles issues des graphes, et les combiner à celles textuelles des nœuds utilisateurs. Dans cette section, nous résumons les contributions et soulignons les apports des travaux de cette thèse. La section 6.1.1, présente l'état de l'art sur l'analyse de la controverse sur le web et les réseaux sociaux en particulier, ainsi que sur la représentation des données sous forme de graphes et les réseaux de neurones graphiques (GNN). La section 6.1.2 porte sur l'explicabilité de la controverse sur Twitter, du point de vue des communautés d'utilisateurs. La section 6.1.3 se focalise sur la détection des posts controversés Reddit, avec une étude sur les sujets Twitter afin de montrer la généralisation possible de l'approche. La section 6.1.4 se concentre sur la quantification de la controverse du point de vue des utilisateurs.

6.1.1 État de l'art

Dans l'état de l'art présenté dans le chapitre 2, nous nous sommes focalisés sur la thématique principale de cette thèse, l'étude de la controverse dans la littérature, ainsi que l'approche d'apprentissage profond la plus exploitée, les réseaux de neurones graphiques (GNN).

Tout d'abord, nous avons présenté des travaux un peu plus anciens qui ont défini la controverse et ont étudié ce phénomène sur le web, à partir de différentes sources de données, comme l'encyclopédie en ligne Wikipédia, des fils de commentaires dans les journaux en ligne, ou encore des réseaux sociaux tels que Twitter ou Reddit. Dans la littérature comme dans cette thèse, nous avons principalement étudié les tâches de détection et de quantification de la controverse.

Dans une seconde partie, nous avons mis en avant les avancées réalisées autour des réseaux de neurones graphiques (GNN). Nous avons présenté les différentes théories autour des GNN, en portant un intérêt particulier sur les réseaux de neurones à convolution graphiques. Nous avons ensuite présenté différentes architectures GNN pour représenter les nœuds d'un graphe comme un vecteur de représentation, ainsi que le graphe. Enfin, différentes méthodes d'apprentissage et tâches appliquées sur les graphes sont présentées. Les travaux des chapitres 4 et 5 se basent sur ces approches, et sont appliqués dans le cas de la controverse. Dans ces chapitres, nous avons produit des figures originales afin de synthétiser les différents travaux de la littérature.

6.1.2 Explication de la controverse par l'analyse des communautés polarisées sur Twitter

Le chapitre 3 présente une méthode originale visant à analyser et expliquer les communautés qui s'expriment sur des sujets controversés sur Twitter. Notre approche est basée sur la méthode SHAP et sur un classifieur de tweets dans les communautés, permettant d'analyser les contributions des caractéristiques sur la prédiction du classifieur. Ces caractéristiques représentent le comportement des usagers dans les communautés via le graphe des interactions et le texte des messages.

Notre approche se base sur 3 principales étapes : (1) la création d'un graphe utilisateur de retweets, (2) la quantification du sujet, combinant un score à partir des données structurelles du graphe avec un autre score basé sur la précision de notre modèle a correctement classifié les tweets dans les bonnes communautés utilisateurs, puis (3) l'analyse des contributions des caractéristiques du contenu des Tweets à la prédiction des communautés d'utilisateurs du modèle. Différentes caractéristiques sont analysées, qu'elles soient textuelles (tokens TF-IDF et BERT) ou conceptuelles (LIWC). Une illustration des visualisations a été effectuée sur deux sujets présentant des mesures de controverse élevées.

Ces travaux ont mené à une publication dans la conférence internationale IDEAS, paru en 2023 [Ben+23c].

6.1.3 Détection des posts controversés à l'aide des réseaux de neurones graphiques

Le chapitre 4 présente une méthode originale afin de détecter les discussions (ou posts) controversés sur Reddit. Dans ce chapitre, nous observons la controverse du point de vue des échanges entre utilisateurs, sur de courtes périodes de temps après la publication des posts, à partir de l'arbre de commentaires. Notre approche se base sur les informations structurelles et textuelles des discussions entre les utilisateurs. Pour ce faire, nous utilisons les réseaux de neurones graphiques (GNN) et réduisons notre approche à une tâche de classification de graphe.

Notre méthode se divise en 4 étapes : (1) la construction du graphe utilisateurs autour de la discussion, (2) le calcul de la représentation des utilisateurs à partir du contenu des tweets, (3) l'apprentissage des nouvelles représentations utilisateurs à l'aide de notre modèle basé sur les GNN et d'une représentation agrégée du graphe et enfin (4) la classification du graphe comme controversé ou non. On montre notamment que dans la tâche de détection des posts controversés, notre approche obtient des résultats très intéressants sur certains jeux de données, ou sont proches de ceux des modèles de références sur d'autres jeux. Afin d'étudier la généralisation de notre approche, nous avons reproduit l'étude sur un autre réseau social (Twitter), utilisant des graphes utilisateurs (graphe de retweets) différents.

Ces travaux ont mené à une publication dans la conférence internationale WISE, paru en 2021 [Ben+21], ainsi qu'un article court à la conférence PFIA 2023 [Ben+23a]. Ces travaux ont aussi abouti à un article étendu dans le journal WWW en 2023 [Ben+23b].

6.1.4 Quantification du point de vue utilisateur de la controverse sur Twitter à l'aide de réseaux de neurones graphiques

Le chapitre 5 présente une méthode originale afin de quantifier la controverse à partir des utilisateurs pour un sujet sur Twitter. Ce chapitre se focalise sur la polarisation des utilisateurs autour de communautés comme définition de la controverse, en étudiant les graphes de retweets comme graphe utilisateur.

Une modélisation théorique est proposée afin de quantifier la controverse, se basant sur l'espérance d'un utilisateur de participer à un sujet controversé. Ce score est ensuite estimé de manière empirique, en se basant sur la probabilité estimée de chaque utilisateur de participer à un sujet controversé. Nous avons démontré que minimiser cet estimateur avec notre fonction de perte revient à maximiser notre score de quantification. Afin d'estimer cette probabilité, notre approche se base sur les GNN et se divise en 3 étapes : (1) la création du graphe utilisateur de retweets et l'initialisation de la représentation des utilisateurs à partir de leurs tweets, (2) le calcul de la nouvelle représentation des utilisateurs, à partir d'un modèle de type GNN à plusieurs couches et (3) la prédiction à partir de ces représentations de l'appartenance des utilisateurs à un sujet controversé. À noter que contrairement au chapitre 4, la nouvelle représentation des utilisateurs est affinée par notre modèle.

Afin d'analyser la qualité de notre métrique, un protocole d'évaluation est proposé, évaluant l'entropie-croisée de nos modèles avec ceux de notre base de référence, sur un jeu de test contenant des sous-graphes centrés sur des utilisateurs à différents niveaux de voisinage k. Nous étudions aussi l'évolution de notre score de controverse selon ce niveau de voisinage k. Au vu des premiers résultats, nous montrons que notre modèle combinant les informations textuelles (tweets) et structurelles (graphe) obtient les meilleures performances. Le score de quantification découlant de ce modèle montre aussi une meilleure séparation des sujets controversés et non controversés que les variantes de ce score utilisant uniquement le texte ou la structure.

6.2 Perspectives

Différentes perspectives de recherche se dégagent en vue d'améliorer plus encore l'explication, la détection et la quantification de la controverse sur les médias sociaux.

6.2.1 Tâches autour de la controverse

Cette section décrit différentes tâches autour de la thématique de l'analyse de la controverse qui semblent intéressantes d'étudier.

6.2.1.1 Analyse temporelle et Prédiction de la croissance de la controverse.

À l'aide de notre approche sur la quantification de la controverse (chapitre 5), nous avons créé un premier score permettant de quantifier un sujet à partir de ces utilisateurs, en utilisant son voisinage proche ainsi que les caractéristiques textuelles des utilisateurs concernés. L'analyse que nous avons menée était globale et ne tient

pas compte de l'évolution temporelle d'une discussion. Or, l'importance de l'aspect temporel dans l'étude des comportements sociaux a déjà été étudié. Mu et al. [Mu+23] montrent l'impact de la dérive temporelle sur la détection de la prise de positions des utilisateurs à l'égard de la vaccination contre la COVID-19 sur Twitter. Concernant la controverse, des travaux très récents ont été produits par Lachi et al. [Lac+23] sur la détection des communautés en analysant leur évolution et leur polarisation par rapport aux utilisateurs anti-vaccination et pro-vaccination au fil du temps. Nous proposons de continuer à explorer notre méthode de quantification de la controverse en incluant cette notion de temporalité. Il s'agit de quantifier ainsi la controverse sur des périodes de temps définies (jour, semaine, etc.). La figure 6.1 montre l'évolution de la controverse avant et après un événement. L'accélération de la controverse peut être aussi étudiée, en se focalisant sur l'augmentation ou diminution soudaine du score à travers le temps.

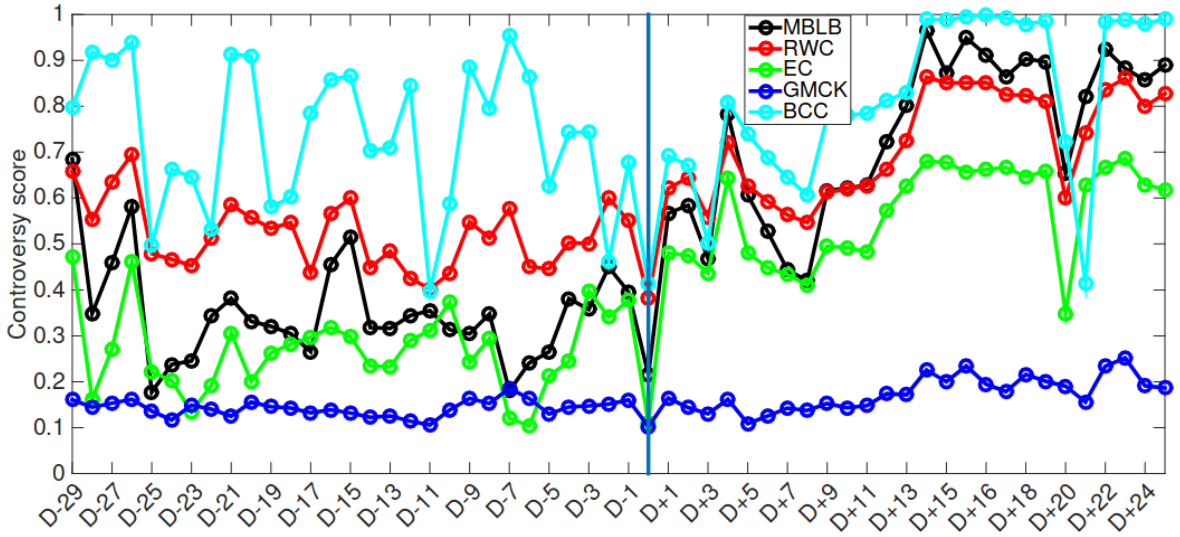

Figure 6.1 – *Score de controverse rwc_score Scores sur 56 graphes de retweets différents concernant des discussions à propos du président Hugo Chavez. Ces scores montrent l'évolution de ce score avant et après l'annonce de la mort du président Hugo Chavez. L'image provient des travaux de Garimella et al.* [*Gar+18a*].

L'analyse des caractéristiques prédictives des communautés participant à un sujet controversé au cours du temps, ainsi que les sentiments associés, peuvent compléter cette étude, comme le propose Ebadi et al. [Eba+21]. De plus, à la suite de nos études sur l'analyse des communautés controversées (chapitre 3), une étude temporelle complémentaire peut permettre de montrer l'évolution des mots ou concepts caractérisant les communautés au cours du temps. Étudier l'évolution de la controverse permettrait de comprendre quels peuvent être les éléments déclencheurs dans les groupes d'utilisateurs d'une controverse.

Pour compléter cette analyse temporelle, il est possible de prédire l'évolution de la controverse et détecter dans des discussions des signaux précurseurs d'une controverse comme le montrent Jamra, Savonnet et Leclercq [JSL22]. Cela peut permettre par exemple aux décideurs dans certains domaines d'anticiper certaines actions à prendre

pour prévenir une controverse ou tenter de la réduire. Nous proposons de continuer à combiner les interactions entre utilisateurs avec le contenu de leurs messages, et d'exploiter les techniques de GNN pour mesurer et prévoir l'évolution du graphe utilisateurs au fil du temps. Ainsi, la combinaison de la dimension spatiale, avec la propagation des informations d'un nœud à ses voisins, avec la dimension temporelle, par l'extraction des motifs séquentiels sur des séries temporelles, permettrait un meilleur suivi de la croissance d'une controverse. Différents problèmes sous-adjacents sont alors à étudier :

- Prédiction du comportement dynamique des utilisateurs : il peut être intéressant de prédire les potentielles actions et intentions de nouveaux utilisateurs [AAS22]. Les acteurs concernés par la controverse peuvent alors proposer du contenu adapté, notamment au début.
- Prévision de l'évolution de la controverse. La prévision de l'évolution de la controverse permettrait aux entités concernées de travailler plus efficacement. Par exemple, si une augmentation est prédite, il est possible de travailler sur la réduction de la controverse, en limitant la circulation de fausses informations [MTP22]. Un exemple de visualisation de ces prévisions sur un sujet est montré dans la figure 6.2.

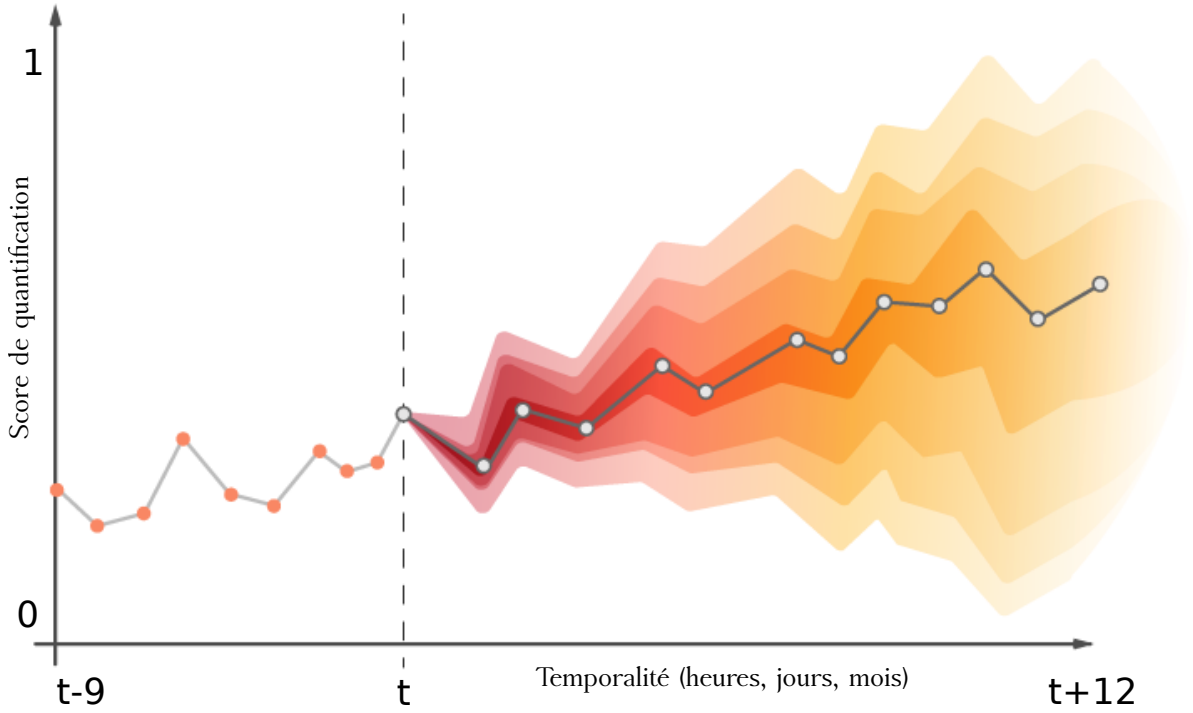

FIGURE 6.2 – *Exemple de visualisation permettant d'analyser les prévisions faites, de l'évolution de la controverse sur les réseaux sociaux. L'axe des abscisses représente l'aspect temporel (jours, mois, trimestres, etc.), et l'axe des ordonnées les prédictions faites sur la métrique voulue. La partie droite de l'axe en pointillé représente les prédictions faites dans le futur, quand la partie gauche représente les valeurs de la métrique dans le passé.*

Des travaux similaires dans les médias sociaux ont été réalisés dans différents

domaines. Dans le monde financier, les méthodes d'apprentissage automatique appliquées sur des données issues des réseaux sociaux sont très souvent utilisées pour prédire le cours des actions [SKR22; Akb+23]. La prévision de cette évolution peut aussi aider à prédire l'issue d'un débat controversé [Mas+23].

6.2.1.2 Génération de la controverse dans les médias sociaux

La génération de sujets controversés, ou a minima l'intervention de modèles d'IA dans des sujets controversés, pourrait cependant changer la manière dont nous traitons l'information.

Selon le contexte et l'objectif visé, la controverse possède plusieurs avantages potentiels. Elle peut stimuler la réflexion en incitant les individus à examiner de manière approfondie un sujet donné, à rechercher des informations et à remettre en question leurs croyances préconçues [Kra+22]. De plus, elle peut favoriser le débat, en encourageant l'expression de différents points de vue. La controverse peut également susciter l'engagement du public, les discussions animées les entourant pouvant captiver l'attention et encourager la participation du public. Elle est un moyen efficace de promouvoir la diversité des opinions, en permettant à une variété de perspectives d'être entendues. La controverse peut contribuer à faire évoluer les normes sociales en mettant en lumière des questions discutables et en suscitant des débats publics, par exemple en catalysant des réformes en mobilisant l'opinion publique en faveur de changements politiques ou sociaux. Enfin, la controverse peut aussi encourager l'innovation en provoquant la remise en question et en incitant les gens à rechercher des solutions alternatives. Toutefois, ces travaux peuvent eux-mêmes être considérés comme controversés. Il est important de les utiliser de manière éthique, en évitant toute manipulation de l'opinion publique, et de reconnaître que certains sujets peuvent parfois entraîner des tensions et des conflits.

Dans ce contexte, l'intervention d'agents conversationnels dans des sujets controversés comme les discussions autour du vaccin pour la Covid-19 vont dans ce sens [Zha+22], notamment pour éviter la prolifération de fausses informations. Deux idées peuvent être mises en place afin de générer de la controverse :

- **Modification de la connectivité entre utilisateurs.** La création de liens entre les utilisateurs, pourrait permettre à des sujets de devenir plus controversés. Les liens entre utilisateurs peuvent créer des échos et des bulles informationnelles ("filter bubble" en anglais) qui renforcent les croyances existantes et rendent plus difficile la remise en question des opinions. Cependant, il est également possible que ces liens favorisent des discussions constructives et des débats équilibrés s'ils encouragent la diversité des opinions et la recherche d'informations objectives. La modification du graphe de discussion est déjà appliquée pour détecter et traiter les rumeurs [He+21]. Elle peut aussi aider à maximiser l'influence de certains utilisateurs centraux [Kum+22].
- **Génération de contenus.** La création et la génération de contenus afin de créer, réduire ou accroitre la controverse, permettrait également d'influencer directement des sujets. Cela correspondrait à la création délibérée de matériel, que ce soit sous forme de textes, d'images, de vidéos ou d'autres médias. Des études

ont montré l'impact que certaines interventions peuvent avoir dans la réduction de l'animosité au sein de certaines communautés controversées [Har+22].

6.2.2 Nouvelles approches autour de l'apprentissage automatique

Afin d'étudier ces perspectives, des approches novatrices basées sur de l'apprentissage automatique et notamment des GNN sont étudiées.

6.2.2.1 GNN et prédiction des liens entre utilisateurs.

Concernant les tâches de prédiction de l'évolution temporelle de la controverse, l'étude de potentielles connexions entre utilisateurs semble être pertinente. Afin d'anticiper l'évolution de la controverse, la prédiction de liens entre les nœuds d'un graphe est une tâche assimilée aux GNN (cf. section 2.3.4.2) et peut constituer une solution. Les modèles GNN classiques adoptent généralement une approche de transmission de messages centrée sur les nœuds, agrégeant les informations des voisins de manière récursive. Ces modèles négligent ainsi les informations topologiques telles que l'emplacement des nœuds et leurs rôles. Cependant, il est important de noter que ces informations topologiques négligées peuvent s'avérer cruciales pour la prédiction des liens. Ai et al. [Ai+22] introduisent une approche novatrice pour prédire des liens, en étiquetant les chemins. Ils capturent les informations topologiques environnantes des nœuds cibles, puis incorporent cette structure dans un modèle GNN. Cette méthode fusionne les structures topologiques et les caractéristiques des nœuds pour tirer pleinement parti des informations du graphe. D'autres travaux traitent les principales limitations des GNN pour prédire les liens, comme leur incapacité à compter les triangles ou à distinguer les nœuds automorphes (nœuds aux rôles structurels identiques) [Cha+22].

Par ailleurs, la prédiction de l'évolution de la controverse, évoquée dans la section 6.2.1.1, reste un problème de prédiction de séries temporelles. On veut exploiter cette évolution à partir de données temporelles multivariées via des GNN. Les GNN exploitent des relations de dépendance prédéfinies, or ces relations ne peuvent être définies d'avance dans le cas de séries temporelles. Des solutions possibles sont proposées pour une meilleure exploitation des GNN sur des séries temporelles multivariées [Wu+20]. Elles ont été appliquées récemment au trafic routier [Zha+18c], au domaine de la finance [Che+21] ou encore à la prévision de l'intensité en carbone des réseaux électriques [ZW23]. Elles permettent essentiellement de capturer, via un mécanisme d'apprentissage, les relations non linéaires qui peuvent exister entre les variables. Le problème de prédiction temporelle ("Time series" en anglais) représente un vrai challenge dans de nombreux domaines. Dans notre cas, l'utilisation de graphes à différentes périodes de temps (par exemple, trois graphes représentant les données à J-2, J-1 et le jour J) pourraient aider à capturer l'évolution du sujet.

De nouvelles architectures GNN, adaptées aux séries temporelles, voient aussi le jour. Wang et Aste [WA22] proposent une architecture de bout en bout pour la prédiction de séries temporelles multivariées, intégrant un GNN spatio-temporel avec un module de filtrage matriciel spécifique. Ce module génère des sous-graphes de corrélation filtrés à partir de séries temporelles multivariées avant de les introduire

dans le GNN. Contrairement aux méthodes existantes adoptées dans les GNN, ce modèle exploite explicitement le filtrage des séries temporelles pour surmonter le faible rapport signal/bruit typique des données des systèmes complexes.

6.2.2.2 GNN et Génération de la controverse.

L'approche générative de la controverse présente un champ très vaste de possibilités. De nouvelles approches d'apprentissage automatique de modèles génératifs ont vu le jour [Fat+22], dont deux majeures sont présentées dans la suite. Premièrement, les réseaux neuronaux générateurs de textes, tels que les modèles basés sur le modèle de langage GPT, sont capables de générer du texte de manière automatisée en apprenant à partir de grandes quantités de données textuelles [Zha+23a]. Les utilisateurs peuvent fournir un début de phrase, et le modèle complète le texte de manière cohérente. Cela peut être utilisé pour générer du contenu controversé à partir de contenu semblable.

Par ailleurs, les modèles adversariaux génératifs (GAN) [Goo+14] sont de plus en plus utilisés, pour produire des contenus synthétiques similaires à des données réelles, à partir de requêtes précises. Ces contenus peuvent être du texte, des images, du son ou des vidéos. Ces réseaux reposent sur un cadre d'entraînement compétitif entre deux réseaux neuronaux distincts : le générateur et le discriminateur. Le générateur apprend à produire des données synthétiques à partir d'un espace latent. Le discriminateur apprend à distinguer les données réelles et les données générées. Ce processus vise à améliorer constamment la capacité du générateur à produire des données de plus en plus réalistes, tout en renforçant la capacité du discriminateur à les discriminer correctement. Les GAN ont été largement utilisés dans des domaines tels que la génération d'images, la synthèse de données et la modification de contenus multimédia. Toutefois, leur potentiel d'utilisation malveillante soulève des préoccupations éthiques et de sécurité. Ce type de modèles est aussi utilisé pour la détection de rumeurs dans les réseaux sociaux [Sun+22]. L'impact de ces données synthétiques générées peut constituer un tournant dans l'analyse de sentiments des communautés controversées [Imr+22].

La génération de graphes à l'aide des GNN peut aussi apporter de la consistance dans la génération de contenu controversé [Wan+21]. Cependant, la génération de graphes reste un problème peu étudié par les modèles adversariaux, et la combinaison de ce type de modèles avec les GNN reste encore à explorer. Des premières méthodes commencent à voir le jour [RWL23], mais leur potentiel impact sur de larges graphes, combinant informations structurelles et textuelles, reste à notre connaissance à étudier.

6.2.3 Application au domaine de la santé

Les différentes approches présentées dans cette thèse peuvent être appliquées à différents types de données et de domaines. Dans ce manuscrit ainsi que dans nos publications, nous avons utilisé des jeux de données collectés sur Twitter et Reddit, issus de la littérature, afin de permettre la reproductibilité de nos approches [Rup+20]. En parallèle des expérimentations réalisées pendant cette thèse, nous avons également récolté des données dans le domaine de la santé. En effet, ces travaux sont financés

par la fondation JANSSEN-HORIZON pour le projet Controverse [1], avec l'objectif à terme de fournir aux professionnels de santé des outils pour analyser, détecter et quantifier la controverse autour de thématiques médicales dans les médias sociaux. Les médias sociaux, permettent une grande liberté d'expression. Ils sont un moyen de communication populaire parmi les patients pour partager leur vécu de la maladie, rechercher facilement des informations et obtenir du soutien. Par exemple, en France, la lecture de certains forums est même recommandée par les principaux organismes impliqués dans la recherche contre le cancer tels que l'INCa et la Ligue Contre le Cancer. Toutes les informations disséminées dans ces médias sociaux peuvent être utilisées comme un vaste réseau de capteurs pour la modélisation de la santé publique à l'échelle de la population [Dic+22]. Par exemple, de nombreux travaux ont exploité les médias sociaux pour analyser la propagation des maladies [Pur+20] ou pour analyser la qualité de vie après un cancer du sein [Joh+22]. Ces travaux fournissent des preuves solides qu'il existe un réel "signal" dans les médias sociaux, qui peut être exploité pour différentes applications liées à la santé. La mise en perspective des controverses dans ces données permet de mettre en lumière des ressentis de patients méconnus afin d'aider les professionnels de santé dans leur travail d'alliance thérapeutique, ou encore d'éviter la prolifération de fausses informations.

Le cas d'étude de ce projet portait initialement sur l'étude des INMs (Intervention non médicamenteuse) dans le cadre du traitement du cancer [MD23]. En effet, ces pratiques, telles que l'activité physique régulière, le recours à un régime nutritionnel adapté, l'utilisation de drogues douces comme le CBD [Del22], etc. sont très souvent proposées par les professionnels comme soins de support en complément des traitements biologiques plus classiques [Zha+23b]. Ces nouvelles pratiques sont très souvent discutées en ligne et génèrent de nombreuses controverses. Suite à la pandémie de 2020, ce cas d'étude a été élargi aux polémiques autour de la Covid-19 qui a eu lieu pendant la première année de cette thèse. Nous avons réalisé de premières expérimentations sur les données de santé récoltées, qui nous ont permis d'envisager des applications prometteuses. Par exemple, nous pourrions étudier les influences mises en jeu dans le processus de décision du patient et qui sont souvent mentionnées dans les médias sociaux [LSC22] lors d'une controverse. Un autre aspect important que nous avons également identifié est le risque de désinformation ou de propagation d'informations médicales erronées [Ska+22], très présent dans le cas de controverse. Il est important de détecter au plus tôt ce type de message pour aider les modérateurs.

Une limite à l'application des méthodes développées dans cette thèse au domaine de la santé est la difficulté à acquérir de grands volumes de données. En effet, les données sur ces thèmes présentes dans Reddit sont limitées et Twitter a modifié son API au milieu de cette thèse, ne permettant plus de large récolte. Il faut donc envisager des méthodes d'augmentation prenant en compte le graphe [Din+22]. Il est également possible d'intégrer des sources externes comme des graphes de connaissances ("Knowledge graphs" en anglais) [LHZ22]. Finalement, il peut être intéressant d'étudier en quoi l'exploitation de telles connaissances peut améliorer l'explication, la détection et la quantification de la controverse. Il s'agit d'un pré-requis pour que les médias sociaux deviennent un support pertinent à une médecine ouverte, participative et collaborative.

1. http://advanse.lirmm.fr/AD/Controverse.php

BIBLIOGRAPHIE

[AAS22] Amina ALMARZOUQI, Ahmad ABURAYYA et Said A SALLOUM. « Prediction of user's intention to use metaverse system in medical education : A hybrid SEM-ML learning approach ». In : *IEEE access* 10 (2022), p. 43421-43434 (cf. page 126).

[Add+17] Aseel ADDAWOOD, Rezvaneh REZAPOUR, Omid ABDAR et Jana DIESNER. « Telling Apart Tweets Associated with Controversial versus Non-Controversial Topics ». In : *Proceedings of the Second Workshop on NLP and Computational Social Science, NLP+CSS@ACL 2017, Vancouver, Canada, August 3, 2017*. Association for Computational Linguistics, 2017, p. 32-41 (cf. page 35).

[Ai+22] Baole AI, Zhou QIN, Wenting SHEN et Yong LI. « Structure enhanced graph neural networks for link prediction ». In : *arXiv preprint arXiv :2201.05293* (2022) (cf. page 128).

[Akb+23] M Eren AKBIYIK, Mert ERKUL, Killian KÄMPF, Vaiva VASILIAUSKAITE et Nino ANTULOV-FANTULIN. « Ask" who", not" what" : Bitcoin volatility forecasting with twitter data ». In : *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 2023, p. 688-696 (cf. page 127).

[Al-+18] Mahmoud AL-AYYOUB, Abdullateef RABAB'AH, Yaser JARARWEH, Mohammed Naji AL-KABI et Brij Bhooshan GUPTA. « Studying the controversy in online crowds' interactions ». In : *Appl. Soft Comput.* 66 (2018), p. 557-563 (cf. page 36).

[Ben+21] Samy BENSLIMANE, Jérôme AZÉ, Sandra BRINGAY, Maximilien SERVAJEAN et Caroline MOLLEVI. « Controversy Detection : a Text and Graph Neural Network Based Approach ». In : *Web Information Systems Engineering–WISE 2021 : 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part I 22*. Springer. 2021, p. 339-354 (cf. pages 85, 123).

[Ben+23a] Samy BENSLIMANE, Jérôme AZÉ, Sandra BRINGAY, Caroline MOLLEVI et Maximilien SERVAJEAN. « Détection de la controverse : une approche basée sur les réseaux de neurones, appliquée aux graphes et aux textes ». In : *CNIA 2023-Conférence Nationale en Intelligence Artificielle*. 2023, p. 50-51 (cf. pages 85, 123).

[Ben+23b] Samy BENSLIMANE, Jérôme AZÉ, Sandra BRINGAY, Maximilien SERVAJEAN et Caroline MOLLEVI. « A text and GNN based controversy detection method on social media ». In : *World Wide Web* 26.2 (2023), p. 799-825 (cf. pages 85, 123).

[Ben+23c] Samy BENSLIMANE, Thomas PAPASTERGIOU, Jérôme AZÉ, Sandra BRINGAY, Caroline MOLLEVI et Maximilien SERVAJEAN. « Explaining controversy through community analysis on Twitter ». In : *Proceedings of the 27th*

International Database Engineered Applications Symposium. 2023, p. 148-155 (cf. pages 62, 123).

[BKV17] Kaspar BEELEN, Evangelos KANOULAS et Bob van de VELDE. « Detecting Controversies in Online News Media ». In : *40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2017, p. 1069-1072 (cf. page 30).

[Boy+22] Ryan BOYD, Ashwini ASHOKKUMAR, Sarah SERAJ et James PENNEBAKER. « The Development and Psychometric Properties of LIWC-22 ». In : (fév. 2022) (cf. pages 16, 35, 63, 66).

[Bro+20] Tom BROWN, Benjamin MANN, Nick RYDER, Melanie SUBBIAH, Jared D KAPLAN, Prafulla DHARIWAL, Arvind NEELAKANTAN, Pranav SHYAM, Girish SASTRY, Amanda ASKELL et al. « Language models are few-shot learners ». In : *Advances in neural information processing systems* 33 (2020), p. 1877-1901 (cf. page 14).

[Bru+14] Joan BRUNA, Wojciech ZAREMBA, Arthur SZLAM et Yann LECUN. *Spectral Networks and Locally Connected Networks on Graphs*. 2014. eprint : 1312.6203 (cs.LG) (cf. page 47).

[Cha+22] Benjamin Paul CHAMBERLAIN, Sergey SHIROBOKOV, Emanuele ROSSI, Fabrizio FRASCA, Thomas MARKOVICH, Nils HAMMERLA, Michael M BRONSTEIN et Max HANSMIRE. « Graph neural networks for link prediction with subgraph sketching ». In : *arXiv preprint arXiv :2209.15486* (2022) (cf. page 128).

[Che+19] Qi CHEN, Yuqiang FENG, Luning LIU et Xianyun TIAN. « Understanding consumers' reactance of online personalized advertising : A new scheme of rational choice from a perspective of negative effects ». In : *International Journal of Information Management* 44 (2019), p. 53-64 (cf. page 11).

[Che+21] Dawei CHENG, Fangzhou YANG, Sheng XIANG et Jin LIU. « Financial Time Series Forecasting with Multi-Modality Graph Neural Network ». In : *Pattern Recognition* 121 (août 2021), p. 108218 (cf. page 128).

[Col+17] Mauro COLETTO, Kiran GARIMELLA, Aristides GIONIS et Claudio LUCCHESE. « Automatic controversy detection in social media : A content-independent motif-based approach ». In : *Online Social Networks and Media* 3-4 (2017), p. 22-31 (cf. pages 39, 40).

[Con+11] Michael D. CONOVER, Jacob RATKIEWICZ, Matthew R. FRANCISCO, Bruno GONÇALVES, Filippo MENCZER et Alessandro FLAMMINI. « Political Polarization on Twitter ». In : *Proceedings of the Fifth International Conference on Weblogs and Social Media*. The AAAI Press, 2011 (cf. page 32).

[DA15] Shiri DORI-HACOHEN et James ALLAN. « Automated Controversy Detection on the Web ». In : *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR*. T. 9022. Lecture Notes in Computer Science. 2015, p. 423-434 (cf. page 28).

[DBV16] Michaël DEFFERRARD, Xavier BRESSON et Pierre VANDERGHEYNST. « Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering ». In : NIPS'16. Curran Associates Inc., 2016, p. 3844-3852 (cf. pages 47-49, 54).

[Del22] Alexis Delaforge. « Visualisation pour l'interprétation et l'explicabilité des prédictions issues de modèles d'apprentissage profond en TAL ». Thèse de doct. Université de Montpellier, 2022 (cf. page 130).

[Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova. « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding ». In : *NAACL-HLT Conference : Human Language Technologies, Volume 1*. 2019, p. 4171-4186 (cf. pages 14, 35, 37, 65, 67, 89, 93, 95).

[DGL13] Luc Devroye, László Györfi et Gábor Lugosi. *A probabilistic theory of pattern recognition*. T. 31. Springer Science & Business Media, 2013 (cf. page 109).

[Dic+22] Grazia Dicuonzo, Graziana Galeone, Matilda Shini et Antonella Massari. « Towards the use of big data in healthcare : A literature review ». In : *Healthcare*. T. 10. 7. MDPI. 2022, p. 1232 (cf. page 130).

[Din+22] Kaize Ding, Zhe Xu, Hanghang Tong et Huan Liu. « Data augmentation for deep graph learning : A survey ». In : *ACM SIGKDD Explorations Newsletter* 24.2 (2022), p. 61-77 (cf. page 130).

[DJA16] Shiri Dori-Hacohen, David D. Jensen et James Allan. « Controversy Detection in Wikipedia Using Collective Classification ». In : *39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2016, p. 797-800 (cf. page 29).

[Eba+21] Ashkan Ebadi, Pengcheng Xi, Stéphane Tremblay, Bruce Spencer, Raman Pall et Alexander Wong. « Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing ». In : *Scientometrics* 126 (2021), p. 725-739 (cf. page 125).

[Ema+20] Hanif Emamgholizadeh, Milad Nourizade, Mir Saman Tajbakhsh, Mahdieh Hashminezhad et Farzaneh Nasr Esfahani. « A framework for quantifying controversy of social network debates using attributed networks : biased random walk (BRW) ». In : *Soc. Netw. Anal. Min.* 10.1 (2020), p. 90 (cf. page 34).

[Fat+22] Noureen Fatima, Ali Shariq Imran, Zenun Kastrati, Sher Muhammad Daudpota et Abdullah Soomro. « A systematic literature review on text generation using deep neural network models ». In : *IEEE Access* 10 (2022), p. 53490-53503 (cf. page 129).

[Gar+16] Kiran Garimella, Michael Mathioudakis, Gianmarco De Francisci Morales et Aristides Gionis. « Exploring controversy in twitter ». In : *Proceedings of the 19th ACM conference on computer supported cooperative work and social computing companion*. 2016, p. 33-36 (cf. page 32).

[Gar+18a] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis et Michael Mathioudakis. « Quantifying Controversy on Social Media ». In : *ACM Trans. Soc. Comput.* 1.1 (2018), 3 :1-3 :27 (cf. pages 16, 19, 32, 34, 36, 43, 58, 65, 67, 81, 100, 101, 105, 106, 108, 118, 125).

[Gar+18b] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis et Michael Mathioudakis. « Reducing Controversy by Connecting Opposing Views ». In : *Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*. 2018, p. 5249-5253 (cf. page 34).

[GG16] Yarin GAL et Zoubin GHAHRAMANI. « Dropout as a bayesian approximation : Representing model uncertainty in deep learning ». In : *international conference on machine learning*. PMLR. 2016, p. 1050-1059 (cf. page 120).

[Gil+17] Justin GILMER, Samuel S SCHOENHOLZ, Patrick F RILEY, Oriol VINYALS et George E DAHL. « Neural message passing for quantum chemistry ». In : *International conference on machine learning*. PMLR. 2017, p. 1263-1272 (cf. pages 47, 51).

[Goo+14] Ian GOODFELLOW, Jean POUGET-ABADIE, Mehdi MIRZA, Bing XU, David WARDE-FARLEY, Sherjil OZAIR, Aaron COURVILLE et Yoshua BENGIO. « Generative adversarial nets ». In : *Advances in neural information processing systems* 27 (2014) (cf. page 129).

[GT22] Biraja GHOSHAL et Allan TUCKER. « On calibrated model uncertainty in deep learning ». In : *arXiv preprint arXiv :2206.07795* (2022) (cf. page 120).

[Gue+13] Pedro Henrique Calais GUERRA, Wagner Meira JR., Claire CARDIE et Robert KLEINBERG. « A Measure of Polarization on Social Media Networks Based on Community Boundaries ». In : *Seventh International Conference on Weblogs and Social Media, ICWSM*. The AAAI Press, 2013 (cf. page 34).

[Guo+15] Jinlong GUO, Yujie LU, Tatsunori MORI et Catherine BLAKE. « Expert-Guided Contrastive Opinion Summarization for Controversial Issues ». In : *Proceedings of the 24th ACM International Conference on World Wide Web*. WWW '15 Companion. 2015, p. 1105-1110 (cf. page 38).

[Guo+17] Chuan GUO, Geoff PLEISS, Yu SUN et Kilian Q WEINBERGER. « On calibration of modern neural networks ». In : *International conference on machine learning*. PMLR. 2017, p. 1321-1330 (cf. page 120).

[Ham+20] Tarek HAMDI, Hamda SLIMI, Ibrahim BOUNHAS et Yahya SLIMANI. « A Hybrid Approach for Fake News Detection in Twitter Based on User Features and Graph Embedding ». In : *Distributed Computing and Internet Technology - 16th International Conference, ICDCIT 2020, Bhubaneswar, India, January 9-12, 2020, Proceedings*. T. 11969. Lecture Notes in Computer Science. Springer, 2020, p. 266-280 (cf. page 43).

[Har+22] Rachel HARTMAN, Will BLAKEY, Jake WOMICK, Chris BAIL, Eli J FINKEL, Hahrie HAN, John SARROUF, Juliana SCHROEDER, Paschal SHEERAN, Jay J VAN BAVEL et al. « Interventions to reduce partisan animosity ». In : *Nature human behaviour* 6.9 (2022), p. 1194-1205 (cf. page 128).

[HBL15] Mikael HENAFF, Joan BRUNA et Yann LECUN. « Deep Convolutional Networks on Graph-Structured Data ». In : *arXiv e-prints* (juin 2015) (cf. page 43).

[He+21] Zhenyu HE, Ce LI, Fan ZHOU et Yi YANG. « Rumor detection on social media with event augmentations ». In : *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, p. 2020-2024 (cf. page 127).

[HL13] Pili HU et Wing Cheong LAU. « A Survey and Taxonomy of Graph Sampling ». In : *arXiv :1308.5865 [cs, math, stat]* (2013) (cf. page 101).

[HL19] Jack HESSEL et Lillian LEE. « Something's Brewing ! Early Prediction of Controversy-causing Posts from Discussion Features ». In : *Conference of the North American Chapter of the Association for Computational Linguistics :*

Human Language Technologies, NAACL-HLT. 2019, p. 1648-1659 (cf. pages 10, 37, 38, 43, 57, 88, 95, 97).

[HY18] William Grant Hatcher et Wei Yu. « A Survey of Deep Learning : Platforms, Applications and Emerging Research Trends ». In : *IEEE Access* 6 (2018), p. 24411-24432 (cf. page 13).

[HYL17] William L. Hamilton, Zhitao Ying et Jure Leskovec. « Inductive Representation Learning on Large Graphs ». In : *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems*. 2017, p. 1024-1034 (cf. pages 19, 43, 47, 49-51, 113).

[Imr+22] Ali Shariq Imran, Ru Yang, Zenun Kastrati, Sher Muhammad Daudpota et Sarang Shaikh. « The impact of synthetic text generation for sentiment analysis using GAN based models ». In : *Egyptian Informatics Journal* 23.3 (2022), p. 547-557 (cf. page 129).

[JA16] Myungha Jang et James Allan. « Improving Automated Controversy Detection on the Web ». In : *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR*. ACM, 2016, p. 865-868 (cf. page 29).

[JA18] Myungha Jang et James Allan. « Explaining Controversy on Social Media via Stance Summarization ». In : *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. ACM, 2018, p. 1221-1224 (cf. page 39).

[Jac+14] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann et Mathieu Bastian. « ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software ». In : *PloS one journal* 9 (juin 2014) (cf. pages 33, 75).

[Jan+16] Myungha Jang, John Foley, Shiri Dori-Hacohen et James Allan. « Probabilistic Approaches to Controversy Detection ». In : *25th ACM International Conference on Information and Knowledge Management, CIKM*. 2016, p. 2069-2072 (cf. page 28).

[JDA17] Myungha Jang, Shiri Dori-Hacohen et James Allan. « Modeling Controversy within Populations ». In : *Proceedings of the SIGIR International Conference on Theory of Information Retrieval, ICTIR*. ACM, 2017, p. 141-149 (cf. page 10).

[JL22] Weiwei Jiang et Jiayun Luo. « Graph neural network for traffic forecasting : A survey ». In : *Expert Systems with Applications* 207 (2022), p. 117921. issn : 0957-4174 (cf. page 43).

[Joh+22] Skyler B Johnson, Matthew Parsons, Tanya Dorff, Meena S Moran, John H Ward, Stacey A Cohen, Wallace Akerley, Jessica Bauman, Joleen Hubbard, Daniel E Spratt et al. « Cancer misinformation and harmful information on Facebook and other social media : a brief report ». In : *JNCI : Journal of the National Cancer Institute* 114.7 (2022), p. 1036-1039 (cf. page 130).

[JSL22] Hiba Abou Jamra, Marinette Savonnet et Éric Leclercq. « Identification of Weak Signals in a Temporal Graph of Social Interactions ». In :

IDEAS'22 : International Database Engineered Applications Symposium, Budapest, Hungary, August 22 - 24, 2022. ACM, 2022, p. 34-42 (cf. page 125).

[KA19] Youngwoo KIM et James ALLAN. « Unsupervised explainable controversy detection from online news ». In : *Advances in Information Retrieval : 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*. Springer. 2019, p. 836-843 (cf. pages 30, 38, 56).

[Kit+07] Aniket KITTUR, Bongwon SUH, Bryan A. PENDLETON et Ed H. CHI. « He says, she says : conflict and coordination in Wikipedia ». In : *Proceedings of the 2007 Conference on Human Factors in Computing Systems, CHI 2007, San Jose, California, USA, April 28 - May 3, 2007*. ACM, 2007, p. 453-462 (cf. pages 26, 30, 56).

[KK95] George KARYPIS et Vipin KUMAR. « METIS – Unstructured Graph Partitioning and Sparse Matrix Ordering System, Version 2.0 ». In : (jan. 1995) (cf. pages 16, 32, 33, 65).

[Kra+22] Elizabeth KRAATZ, Jacqueline von SPIEGEL, Robin SAYERS et Anna C BRADY. « Should we "just stick to the facts" ? The benefit of controversial conversations in classrooms ». In : *Theory Into Practice* 61.3 (2022), p. 312-324 (cf. page 127).

[Kul+19] Meelis KULL, Miquel PERELLO NIETO, Markus KÄNGSEPP, Telmo SILVA FILHO, Hao SONG et Peter FLACH. « Beyond temperature scaling : Obtaining well-calibrated multi-class probabilities with dirichlet calibration ». In : *Advances in neural information processing systems* 32 (2019) (cf. page 120).

[Kum+22] Sanjay KUMAR, Abhishek MALLIK, Anavi KHETARPAL et BS PANDA. « Influence maximization in social networks using graph embedding and graph neural network ». In : *Information Sciences* 607 (2022), p. 1617-1636 (cf. page 127).

[KW16] Thomas N KIPF et Max WELLING. « Variational graph auto-encoders ». In : *arXiv preprint arXiv :1611.07308* (2016) (cf. page 56).

[KW17] Thomas N. KIPF et Max WELLING. « Semi-Supervised Classification with Graph Convolutional Networks ». In : *5th International Conference on Learning Representations, ICLR*. OpenReview.net, 2017 (cf. pages 17, 43, 45, 47-49, 51, 54, 87, 90, 92, 96).

[KWH21] Philipp KONCAR, Simon WALK et Denis HELIC. « Analysis and Prediction of Multilingual Controversy on Reddit ». In : *13th ACM Web Science Conference 2021*. 2021, p. 215-224 (cf. pages 38, 39, 63).

[Lac+23] Veronica LACHI, Giovanna Maria DIMITRI, Alessandro DI STEFANO, Pietro LIÒ, Monica BIANCHINI et Chiara MOCENNI. « Impact of the Covid 19 outbreaks on the italian twitter vaccination debat : a network based analysis ». In : *arXiv preprint arXiv :2306.02838* (2023) (cf. page 125).

[Lec+98] Y. LECUN, L. BOTTOU, Y. BENGIO et P. HAFFNER. « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86.11 (1998), p. 2278-2324 (cf. page 44).

[LHZ22] Michelle M LI, Kexin HUANG et Marinka ZITNIK. « Graph representation learning in biomedicine and healthcare ». In : *Nature Biomedical Engineering* 6.12 (2022), p. 1353-1369 (cf. page 130).

[Li+15] Yujia Li, Daniel Tarlow, Marc Brockschmidt et Richard Zemel. « Gated graph sequence neural networks ». In : *arXiv preprint arXiv :1511.05493* (2015) (cf. page 52).

[LL17] Scott M Lundberg et Su-In Lee. « A Unified Approach to Interpreting Model Predictions ». In : *Advances in Neural Information Processing Systems*. Sous la dir. d'I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett. T. 30. Curran Associates, Inc., 2017 (cf. pages 63, 68, 69).

[Lor20] Titouan Lorieul. « Uncertainty in predictions of Deep Learning models for fine-grained classification ». Thèse de doct. Déc. 2020 (cf. pages 110, 111).

[LRK18] John Boaz Lee, Ryan Rossi et Xiangnan Kong. « Graph classification using structural attention ». In : *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, p. 1666-1674 (cf. page 52).

[LSC22] Mingda Li, Jinhe Shi et Yi Chen. « Identifying Influences in Patient Decision-making Processes in Online Health Communities : Data Science Approach ». In : *Journal of Medical Internet Research* 24.8 (2022), e30634 (cf. page 130).

[Lu+23] Jiang Lu, Pinghua Gong, Jieping Ye, Jianwei Zhang et Changshui Zhang. *A Survey on Machine Learning from Few Samples*. 2023. arXiv : 2009.02653 [cs.LG] (cf. page 13).

[Mas+23] Riduan Masud, Muhammad Syamsurrijal, Tawakkal Baharuddin et Muhammad Azizurrohman. « Forecasting political parties and candidates for Indonesia's presidential election in 2024 using twitter ». In : (2023) (cf. page 127).

[MD23] Rami Manochakian et Don S Dizon. « Using social media for patient-driven cancer research ». In : *Nature Reviews Cancer* 23.1 (2023), p. 1-2 (cf. page 130).

[Mik+13] Tomas Mikolov, Kai Chen, Greg Corrado et Jeffrey Dean. « Efficient estimation of word representations in vector space ». In : *arXiv preprint arXiv :1301.3781* (2013) (cf. pages 14, 30, 56).

[Mor+15] Alfredo Jose Morales, Javier Borondo, Juan Carlos Losada et Rosa M. Benito. « Measuring Political Polarization : Twitter shows the two sides of Venezuela ». In : *CoRR* (2015) (cf. pages 32, 39, 40, 108).

[Mou22] Lampros Mouselimis. *fastText : Efficient Learning of Word Representations and Sentence Classification using R*. R package version 1.0.3. 2022 (cf. page 14).

[MPS20] Marcelo Mendoza, Denis Parra et Álvaro Soto. « GENE : Graph generation conditioned on named entities for polarity and controversy detection in social media ». In : *Inf. Process. Manag.* 57.6 (2020), p. 102366 (cf. page 36).

[MTP22] Asutosh Mohapatra, Nithin Thota et P Prakasam. « Fake news detection and classification using hybrid BiLSTM and self-attention model ». In : *Multimedia Tools and Applications* 81.13 (2022), p. 18503-18519 (cf. page 126).

[Mu+23] Yida Mu, Mali Jin, Kalina Bontcheva et Xingyi Song. « Examining Temporalities on Stance Detection Towards COVID-19 Vaccination ». In : *arXiv preprint arXiv :2304.04806* (2023) (cf. page 125).

[Nak+22] Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Mucahid Kutlu, Wajdi Zaghouani, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak et Alex Nikolov. « Overview of the CLEF-2022 CheckThat ! Lab Task 1 on Identifying Relevant Claims in Tweets ». In : (2022) (cf. page 63).

[NVN20] Dat Quoc Nguyen, Thanh Vu et Anh Nguyen. « BERTweet : A pre-trained language model for English Tweets ». In : jan. 2020, p. 9-14 (cf. page 104).

[Pap+19] Raghavendra Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel et Najim Dehak. « Hierarchical Transformers for Long Document Classification ». In : déc. 2019, p. 838-844 (cf. page 98).

[PAS14] Bryan Perozzi, Rami Al-Rfou et Steven Skiena. « Deepwalk : Online learning of social representations ». In : *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, p. 701-710 (cf. pages 43, 55).

[PGA19] Daniel Preoţiuc-Pietro, Mihaela Gaman et Nikolaos Aletras. « Automatically Identifying Complaints in Social Media ». In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, juill. 2019, p. 5008-5019 (cf. page 63).

[PSM14] Jeffrey Pennington, Richard Socher et Christopher Manning. « GloVe : Global Vectors for Word Representation ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, oct. 2014, p. 1532-1543 (cf. page 14).

[Pur+20] Neha Puri, Eric A Coomes, Hourmazd Haghbayan et Keith Gunaratne. « Social media and vaccine hesitancy : new updates for the era of COVID-19 and globalized infectious diseases ». In : *Human vaccines & immunotherapeutics* 16.11 (2020), p. 2586-2593 (cf. page 130).

[Ras+21] Ammar Rashed, Mucahid Kutlu, Kareem Darwish, Tamer Elsayed et Cansın Bayrak. « Embeddings-based clustering for target specific stances : The case of a polarized turkey ». In : *Proceedings of the International AAAI Conference on Web and Social Media*. T. 15. 2021, p. 537-548 (cf. pages 10, 34-36).

[RHH18] Nils Rethmeier, Marc Hübner et Leonhard Hennig. « Learning comment controversy prediction in web discussions using incidentally supervised multi-task CNNs ». In : *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2018, p. 316-321 (cf. page 31).

[RHW+85] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams et al. *Learning internal representations by error propagation*. 1985 (cf. page 62).

[Ruf10] Kaspar Rufibach. « Use of Brier score to assess binary predictions ». In : *Journal of clinical epidemiology* 63.8 (2010), p. 938-939 (cf. page 120).

[Rup+20] Lukas Rupprecht, James C. Davis, Constantine Arnold, Yaniv Gur et Deepavali Bhagwat. « Improving Reproducibility of Data Science Pipelines through Transparent Provenance Capture ». In : *Proc. VLDB Endow.* 13.12 (août 2020), p. 3354-3368. issn : 2150-8097. doi : 10.14778/3415478.3415556 (cf. page 129).

[RWL23] Can Rong, Huandong Wang et Yong Li. « Origin-Destination Network Generation via Gravity-Guided GAN ». In : *arXiv preprint arXiv :2306.03390* (2023) (cf. page 129).

[Sch+18] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov et Max Welling. « Modeling relational data with graph convolutional networks ». In : *The Semantic Web : 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer. 2018, p. 593-607 (cf. pages 47, 51).

[Sha52] Lloyd S. Shapley. *A Value for N-Person Games*. Santa Monica, CA : RAND Corporation, 1952 (cf. page 63).

[SK17] Martin Simonovsky et Nikos Komodakis. « Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs ». In : *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*. Juill. 2017 (cf. page 106).

[Ska+22] Ingjerd Skafle, Anders Nordahl-Hansen, Daniel S Quintana, Rolf Wynn et Elia Gabarron. « Misinformation about COVID-19 vaccines on social media : rapid review ». In : *Journal of medical Internet research* 24.8 (2022), e37367 (cf. page 130).

[SKR22] T Swathi, N Kasiviswanath et A Ananda Rao. « An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis ». In : *Applied Intelligence* 52.12 (2022), p. 13675-13688 (cf. page 127).

[SPP17] Allaparthi Sriteja, Prakhar Pandey et Vikram Pudi. « Controversy Detection Using Reactions on Social Media ». In : *IEEE International Conference on Data Mining Workshops, ICDM Workshops*. IEEE Computer Society, 2017, p. 884-889 (cf. pages 30, 56).

[Sun+19] Chi Sun, Xipeng Qiu, Yige Xu et Xuanjing Huang. « How to Fine-Tune BERT for Text Classification ? » In : *Chinese Computational Linguistics - 18th China National Conference, CCL*. T. 11856. Lecture Notes in Computer Science. Springer, 2019, p. 194-206 (cf. pages 98, 99).

[Sun+22] Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li et Qiaoming Zhu. « Rumor detection on social media with graph adversarial contrastive learning ». In : *Proceedings of the ACM Web Conference 2022*. 2022, p. 2789-2797 (cf. page 129).

[Szn+19] Benjamin Sznajder, Ariel Gera, Yonatan Bilu, Dafna Sheinwald, Ella Rabinovich, Ranit Aharonov, David Konopnicki et Noam Slonim. « Controversy in Context ». In : *CoRR* (2019) (cf. page 29).

[Tix+19] Antoine J-P Tixier, Giannis Nikolentzos, Polykarpos Meladianos et Michalis Vazirgiannis. « Graph classification with 2d convolutional neural networks ». In : *Artificial Neural Networks and Machine Learning–ICANN 2019 : Workshop and Special Sessions : 28th International Conference on Artifi-*

cial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings 28. Springer. 2019, p. 578-593 (cf. page 53).

[Tre+23] Marcos TREVISO, Ji-Ung LEE, Tianchu JI, Betty van AKEN, Qingqing CAO, Manuel R. CIOSICI, Michael HASSID, Kenneth HEAFIELD, Sara HOOKER, Colin RAFFEL, Pedro H. MARTINS, André F. T. MARTINS, Jessica Zosa FORDE, Peter MILDER, Edwin SIMPSON, Noam SLONIM, Jesse DODGE, Emma STRUBELL, Niranjan BALASUBRAMANIAN, Leon DERCZYNSKI, Iryna GUREVYCH et Roy SCHWARTZ. *Efficient Methods for Natural Language Processing : A Survey*. 2023. arXiv : 2209.00099 [cs.CL] (cf. page 14).

[Vas+17] Ashish VASWANI, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N GOMEZ, Łukasz KAISER et Illia POLOSUKHIN. « Attention is all you need ». In : *Advances in neural information processing systems* 30 (2017) (cf. pages 50, 93).

[VBK15] Oriol VINYALS, Samy BENGIO et Manjunath KUDLUR. « Order matters : Sequence to sequence for sets ». In : *arXiv preprint arXiv :1511.06391* (2015) (cf. page 52).

[Vel+18] Petar VELICKOVIC, Guillem CUCURULL, Arantxa CASANOVA, Adriana ROMERO, Pietro LIÒ et Yoshua BENGIO. « Graph Attention Networks ». In : *6th International Conference on Learning Representations, ICLR*. OpenReview.net, 2018 (cf. pages 19, 43, 47, 50-52, 54, 90, 96, 113).

[VRA18] Soroush VOSOUGHI, Deb ROY et Sinan ARAL. « The spread of true and false news online ». In : *science* 359.6380 (2018), p. 1146-1151 (cf. page 11).

[Vuo+08] Ba-Quy VUONG, Ee-Peng LIM, Aixin SUN, Minh-Tam LE, Hady Wirawan LAUW et Kuiyu CHANG. « On ranking controversies in wikipedia : models and evaluation ». In : *Proceedings of the 2008 international conference on Web search and data mining*. 2008, p. 171-182 (cf. pages 26, 56).

[WA22] Yuanrong WANG et Tomaso ASTE. « Sparsification and Filtering for Spatial-temporal GNN in Multivariate Time-series ». In : *arXiv preprint arXiv :2203.03991* (2022) (cf. page 128).

[Wan+21] Sheng WAN, Shirui PAN, Jian YANG et Chen GONG. « Contrastive and generative graph convolutional networks for graph-based semi-supervised learning ». In : *Proceedings of the AAAI conference on artificial intelligence*. T. 35. 11. 2021, p. 10049-10057 (cf. page 129).

[WC16] Lu WANG et Claire CARDIE. « A piece of my mind : A sentiment analysis approach for online dispute detection ». In : *arXiv preprint arXiv :1606.05704* (2016) (cf. pages 29, 30).

[WHD+18] Shiyao WANG, Minlie HUANG, Zhidong DENG et al. « Densely connected CNN with multi-scale feature attention for text classification. » In : *IJCAI*. T. 18. 2018, p. 4468-4474 (cf. page 44).

[Wu+17] Libing WU, Yubo ZHANG, Yong XIE, Abdulhameed ALELAIWI et Jian SHEN. « An Efficient and Secure Identity-Based Authentication and Key Agreement Protocol with User Anonymity for Mobile Devices ». In : *Wirel. Pers. Commun.* 94.4 (2017), p. 3371-3387 (cf. page 10).

[Wu+20] Zonghan WU, Shirui PAN, Guodong LONG, Jing JIANG, Xiaojun CHANG et Chengqi ZHANG. « Connecting the Dots : Multivariate Time Series Forecasting with Graph Neural Networks ». In : *KDD '20 : The 26th ACM*

SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020. ACM, 2020, p. 753-763 (cf. page 128).

[Wu+21] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang et Philip S. Yu. « A Comprehensive Survey on Graph Neural Networks ». In : *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (jan. 2021), p. 4-24. ISSN : 2162-237X, 2162-2388 (cf. pages 14, 19, 43, 44, 47, 52, 54).

[Xie+22] Yaochen Xie, Zhao Xu, Jingtun Zhang, Zhengyang Wang et Shuiwang Ji. « Self-Supervised Learning of Graph Neural Networks : A Unified Review ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022). Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 1-1. ISSN : 1939-3539. DOI : 10.1109/TPAMI.2022.3170559 (cf. pages 43, 55).

[Xu+19] Keyulu Xu, Weihua Hu, Jure Leskovec et Stefanie Jegelka. « How Powerful are Graph Neural Networks ? » In : *arXiv :1810.00826 [cs, stat]* (fév. 2019) (cf. pages 47, 49, 50, 113).

[Yas+12] Taha Yasseri, Robert Sumi, András Rung, András Kornai et János Kertész. « Dynamics of conflicts in Wikipedia ». In : *PloS one* 7.6 (2012), e38869 (cf. pages 28, 30).

[Yin+18] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton et Jure Leskovec. « Hierarchical graph representation learning with differentiable pooling ». In : *Advances in neural information processing systems* 31 (2018) (cf. pages 17, 43, 53, 86, 87, 91, 96).

[Zar+20] Juan Manuel Ortiz de Zarate, Marco Di Giovanni, Esteban Zindel Feuerstein et Marco Brambilla. « Measuring Controversy in Social Networks Through NLP ». In : *27th International Symposium on String Processing and Information Retrieval, SPIRE, Orlando, USA, October 13-15, 2020*. T. 12303. Lecture Notes in Computer Science. 2020, p. 194-209 (cf. pages 15, 35, 36, 59).

[ZF20] Juan Manuel Ortiz De Zarate et Esteban Feuerstein. « Vocabulary-Based Method for Quantifying Controversy in Social Media ». In : *Ontologies and Concepts in Mind and Machine - 25th International Conference on Conceptual Structures, ICCS*. T. 12277. Lecture Notes in Computer Science. Springer, 2020, p. 161-176 (cf. page 36).

[Zha+18a] Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King et Dit-Yan Yeung. « Gaan : Gated attention networks for learning on large and spatiotemporal graphs ». In : *arXiv preprint arXiv :1803.07294* (2018) (cf. pages 47, 50).

[Zha+18b] Muhan Zhang, Zhicheng Cui, Marion Neumann et Yixin Chen. « An end-to-end deep learning architecture for graph classification ». In : *Proceedings of the AAAI conference on artificial intelligence*. T. 32. 1. 2018 (cf. page 52).

[Zha+18c] Ling Zhao, Yujiao Song, Min Deng et Haifeng Li. « Temporal Graph Convolutional Network for Urban Traffic Flow Prediction Method ». In : *CoRR* abs/1811.05320 (2018) (cf. page 128).

[Zha+22] Menghan Zhang, Xue Qi, Ze Chen et Jun Liu. « SocIal bots' involvement in the covid-19 vaccine discussions on Twitter ». In : *International Journal*

of Environmental Research and Public Health 19.3 (2022), p. 1651 (cf. page 127).

[Zha+23a] Chaoning ZHANG, Chenshuang ZHANG, Sheng ZHENG, Yu QIAO, Chenghao LI, Mengchun ZHANG, Sumit Kumar DAM, Chu Myaet THWAL, Ye Lin TUN, Le Luang HUY et al. « A complete survey on generative ai (aigc) : Is chatgpt from gpt-4 to gpt-5 all you need ? » In : *arXiv preprint arXiv :2303.11717* (2023) (cf. page 129).

[Zha+23b] Zhiying ZHAO, Peng LIU, Jing JIN et Wenyan WANG. « Effects of non-drug interventions on anxiety and depression in patients with heart failure : A systematic review based on Bayesian network meta-analysis ». In : *Journal of Psychiatric Research* (2023) (cf. page 130).

[Zho+20] Lei ZHONG, Juan CAO, Qiang SHENG, Junbo GUO et Ziang WANG. « Integrating Semantic and Structural Information with Graph Convolutional Network for Controversy Detection ». In : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*. Association for Computational Linguistics, 2020, p. 515-526 (cf. pages 18, 37, 57, 86, 88, 96, 97).

[ZM18] Chenyi ZHUANG et Qiang MA. « Dual Graph Convolutional Networks for Graph-Based Semi-Supervised Classification ». In : *Proceedings of the 2018 World Wide Web Conference*. WWW '18. Republic et Canton of Geneva, CHE : International World Wide Web Conferences Steering Committee, 2018, p. 499-508 (cf. pages 47, 48).

[ZW23] Xiaoyang ZHANG et Dan WANG. « A GNN-based Day Ahead Carbon Intensity Forecasting Model for Cross-Border Power Grids ». In : *Proceedings of the 14th ACM International Conference on Future Energy Systems*. 2023, p. 361-373 (cf. page 128).

[ZX20] Shuo ZHANG et Lei XIE. « Improving attention mechanism in graph neural networks via cardinality preservation ». In : *IJCAI : Proceedings of the Conference*. T. 2020. NIH Public Access. 2020, p. 1395 (cf. pages 43, 52, 87, 93, 96).

MATÉRIEL SUPPLÉMENTAIRE

Notation et formalisation

TABLE 1 – *Tableau de correspondance des notations utilisées entre nos travaux et ceux présentés dans notre état de l'art (section 2.3) pour la représentation des graphes. Un "−" signifie que la variable n'a pas été introduite dans l'état de l'art.*

Notation		Description
État de l'art	Chapitres 4 et 5	
$-$	t	Sujet Twitter
$-$	p	Post (fil de discussion) Reddit
$-$	s	Subreddit
$-$	G	Ensemble de graphes
G	G_i	Un graphe (identifié par sa position i)
$-$	$g_u^{(k)}$	Sous-graphe centré sur u à k niveau de voisinage
$-$	l_i	Label du $i^{ème}$ graphe
v	u_i	Un nœud utilisateur (identifié par sa position i)
e_{ij}	(u_i, u_j)	Un lien entre les utilisateurs u_i et u_j
$\mathcal{N}(v)$	$\mathcal{N}(u_i)$	Ensemble de voisins du nœud u_i
$\tilde{\mathcal{N}}(v)$	$\mathcal{N}(u_i)$	Ensemble de voisins du nœud u_i, y compris u_i
V	U	Ensemble de nœuds
	E	Ensemble de liens
	A	Matrice d'adjacence
	D	Matrice des degrés de la matrice A
	n	Nombre de nœuds, $n = \|V\|$
	m	Nombre de liens, $m = \|E\|$
	d	Dimension du vecteur de caractéristiques des nœuds
	b	Dimension de la couche cachée d'un réseau de neurones
	$X \in \mathbb{R}^{n \times d}$	Matrice des caractéristiques du graphe d'entrée
	$H^l \in \mathbb{R}^{n \times b}$	Matrice des caractéristiques des nœuds du graphe à la couche l
	$h_v^l \in \mathbb{R}^b$	Représentation du nœud v à la couche l
	W^l	Matrice de poids entraînable de la couche cachée l
	l	Index de la couche cachée du réseau
	σ	Fonction d'activation de la couche du réseau de neurone