Détection d'utilisateurs violents et de menaces dans les réseaux sociaux

Nour El Houda Ben Chaabene

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAS001

Thèse de doctorat

Détection d'utilisateurs violents et de menaces dans les réseaux sociaux

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à TELECOM SudParis en cotutelle avec l'Ecole Nationale des Sciences de l'Informatique de la Manouba

École doctorale n°ED 626 de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Manouba, le 31/01/2022, par

NOUR EL HOUDA BEN CHAABENE

Composition du Jury :

Naoufel Kraiem Professeur, Institut Supérieur de l'Informatique, Tunisie	Président
Amel Borgi Professeur, Institut Supérieur de l'Informatique, Tunisie	Rapporteur
Djamal Benslimane Professeur, Université Claude Bernard Lyon 1, France	Rapporteur
Bruno DEFUDE Professeur, Télécom SudParis, Institut Polytechnique de Paris, France	Examinateur
André PENINOU Maître de conférences, IRIT - Université Paul Sabatier, Toulouse, France	Examinateur
Amel Bouzeghoub Professeur, Télécom SudParis, Institut Polytechnique de Paris France	Directeur de thèse
Henda Ben Ghezala Professeur, Ecole Nationale des Science de l'Informatique, Tunisie	Directeur de thèse
Ramzi Guetari Maitre assistant, Ecole Polytechnique de Tunisie, Tunisie	Invité

A mes chers parents,

à mon adorable mari,

à mon petit bout de chou,

à mes soeurs bien-aimées,

à ma grande famille aimante,

à mes vrais amis,

pour leur amour et soutien.

Remerciements

Avant tout, je remercie Dieu de m'avoir protégé et de m'avoir donné la force et la patience pour finaliser ce travail de recherche.

La réalisation de cette thèse de doctorat a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je commence par le remerciement de mes deux directrices de thèse, **Pr. Amel Bouzeghoub** et **Pr. Henda Ben Ghezala** pour cette extraordinaire opportunité qui m'ont accordé. Je suis extrêmement reconnaissante, mesdames, à vos encouragements, vos suggestions et vos orientations. Merci infiniment pour votre temps précieux, votre collaboration et votre attention qui ont permis à ce travail de voir la lumière.

Je tiens à remercier spécialement mon co-encadrant **Dr. Ramzi Guetari**, qui fut le premier à me faire découvrir le domaine de la recherche. Je désire exprimer, monsieur, mes remerciements les plus sincères pour votre générosité, votre motivation inspirante et votre patience tout au long ces cinq dernières années. Un grand merci encore une fois de votre exigence qui m'a énormément stimulé, sans elle ce travail n'aurait jamais vu le jour tel qu'il est aujourd'hui.

Je voudrais profiter de cette occassion pour exprimer mes plus profonds remerciements à toute l'équipe de Télécom SudParis pour leur accueil chaleureux. En particulier, je remercie **Dr. Badran Raddaoui**, **Mme. Brigitte Houassine** et **Mme. Véronique Guy**. Je tiens également à témoigner toute ma reconnaissance à tous les membres de l'Ecole Nationale des Sciences de l'Informatique. Je

remercie spécialement **Mr. Nabil Morgham** et **Mme. Saida Majdoub** pour leur coopération et leur aide surtout dans les moments difficiles.

J'adresse toutes les expressions de remerciement à mes chers amis et collègues qui m'ont apporté leur soutien moral et intellectuel tout au long de ma démarche. Je remercie particulièrement **Ibtihel**, **Hamza**, **Maryem**, **Khaoula**, **Khouloud**, **Wejdene**, **Maha**, **Ghada**, **Hatem**, **Hafedh** et **Walid**.

Un grand merci à mon père **Mustapha** et ma mère **Zohra**, pour leur amour, leurs conseils, leurs prières ainsi que leur soutien inconditionnel, à la fois moral et financier, qui m'a permis de réaliser les études que je voulais et par conséquent cette thèse de doctorat. Mes remerciemenst vont aussi à mes soeurs **Chahrazed**, **Manel** et **Ines** et mes beaux-frères **Khaled**, **Zaher** et **Ahdi** pour leur appui continu et leurs mots encourageants. Je n'oublirais surtout pas mes nièces et mes neveux **Ameni**, **Arij**, **Mohamed**, **Loujain**, **Alyamen** et **Aline**.

Je remercie mon cher mari **Hassen** pour son soutien quotidien indéfectible et son enthousiasme contagieux à l'égard de mes travaux comme de la vie en général. Notre couple a grandi en même temps que mon projet scientifique, le premier servant de socle solide à l'épanouissement du second.

Ces remerciements ne peuvent s'achever, sans une pensée pour mon coeur et mon adorable petit prince **Mohamed Houssem**. Mon ange, tu es l'un des cadeaux les plus doux que la vie puisse m'apporter, tu es ma source de force. J'espère que tu sera fier de moi comme je le suis.

Résumé

Les réseaux sociaux en ligne font partie intégrante de l'activité sociale quotidienne des gens. Ils fournissent des plateformes permettant de mettre en relation des personnes du monde entier et de partager leurs intérêts. Des statistiques récentes indiquent que 56% de la population mondiale utilisent ces médias sociaux. Cependant, ces services de réseau ont également eu de nombreux impacts négatifs et l'existence de phénomènes d'agressivité et d'intimidation dans ces espaces est inévitable et doit donc être abordée. L'exploration de la structure complexe des réseaux sociaux pour détecter les comportements violents et les menaces est un défi pour l'exploration de données, l'apprentissage automatique et l'intelligence artificielle.

Dans ce travail de thèse, nous visons à proposer de nouvelles approches de détection des comportements violents dans les réseaux sociaux. Nos approches tentent de résoudre cette problèmatique pour plusieurs raisons pratiques. Premièrement, des personnes différentes ont des façons différentes d'exprimer le même comportement violent. Il est souhaitable de conçevoir une approche qui fonctionne pour tout le monde en raison de la variété des comportements et des diverses manières dont ils sont exprimés. Deuxièmement, les approches doivent avoir un moyen de détecter les comportements anormaux potentiels non vus et de les ajouter automatiquement à l'ensemble d'apprentissage. Troisièmement, la multimodalité et la multidimensionnalité des données disponibles sur les sites de réseaux sociaux doivent être prises en compte pour le développement de solutions d'exploration de données qui seront capables d'extraire des informations pertinentes utiles à la détection de comportements violents. Enfin, les approches doivent considérer la nature variable dans le temps des réseaux pour traiter les nouveaux utilisateurs et liens et mettre automatiquement à jour les modèles construits.

A la lumière de cela et pour atteindre les objectifs susmentionnés, les principales contributions de cette thèse sont les suivantes :

— La première contribution propose un modèle de détection des comportements violents sur Twitter. Ce modèle prend en charge la nature dynamique du réseau et est capable d'extraire et d'analyser de données hétérogènes.

— La deuxième contribution introduit une approche de détection des comportements atypiques sur un réseau multidimensionnel. Cette approche se base sur l'exploration et l'analyse des relations entre les individus présents sur cette structure sociale multidimensionnelle.

— La troisième contribution présente un framework d'identification des personnes anormales. Ce cadre intelligent s'appuie sur l'exploitation d'un modèle multidimensionnel qui prend en entrée des données multimodales provenant de plusieurs sources, capable d'enrichir automatiquement l'ensemble d'apprentissage par les comportements violents détectés et considère la dynamicité des données afin de détecter les nouveaux comportements violents qui apparaissent sur le réseau.

Cette thèse décrit des réalisations combinant les techniques d'exploration de données avec les nouvelles techniques d'apprentissage automatique. Pour prouver la performance de nos résultats d'expérimentation, nous nous sommes basés sur des données réelles extraites de trois réseaux sociaux populaires.

Mots-clés : Détection d'anomalie, Analyse des réseaux sociaux complexes, Données multimodales, Réseau multidimensionnel, Détection de communautés

Abstract

Online social networks are an integral part of people's daily social activity. They provide platforms to connect people from all over the world and share their interests. Recent statistics indicate that 56% of the world's population use these social media. However, these network services have also had many negative impacts and the existence of phenomena of aggression and intimidation in these spaces is inevitable and must therefore be addressed. Exploring the complex structure of social networks to detect violent behavior and threats is a challenge for data mining, machine learning, and artificial intelligence.

In this thesis work, we aim to propose new approaches for the detection of violent behavior in social networks. Our approaches attempt to resolve this problem for several practical reasons. First, different people have different ways of expressing the same violent behavior. It is desirable to design an approach that works for everyone because of the variety of behaviors and the various ways in which they are expressed. Second, the approaches must have a way to detect potential unseen abnormal behaviors and automatically add them to the training set. Third, the multimodality and multidimensionality of the data available on social networking sites must be taken into account for the development of data mining solutions that will be able to extract relevant information useful for the detection of violent behavior. Finally, approaches must consider the time-varying nature of networks to process new users and links and automatically update built models.

In the light of this and to achieve the aforementioned objectives, the main contributions of this thesis are as follows :

— The first contribution proposes a model for detecting violent behavior on Twitter. This model supports the dynamic nature of the network and is capable of extracting and analyzing heterogeneous data.

— The second contribution introduces an approach for detecting atypical behaviors on a multidimensional network. This approach is based on the exploration and analysis of the relationships between the individuals present on this multidimensional social structure.

— The third contribution presents a framework for identifying abnormal people. This intelligent framework is based on the exploitation of a multidimensional model which takes as input multimodal data coming from several sources, capable of automatically enriching the learning set by the violent behaviors detected and considers the dynamicity of the data in order to to detect new violent behaviors that appear on the network.

This thesis describes achievements combining data mining techniques with new machine learning techniques. To prove the performance of our experimental results, we sums based on real data taken from three popular social networks.

Keywords : Anomaly detection, Complex social networks analysis, Multimodal data, Multidimensionnal network, Community detection

Acronymes

BIC Bayesian Information Criterion

BPNN Back Propagation Neural Network

CBOW Continuous Bag Of Words

CF Collaborative Filtering

Char-LSTM Char-Long Short-Term Memory

CNN Convolutional Neural Network

CNN-LSTM Convolutional Neural Network - Long Short-Term Memory

CUHK Chinese Univeristy of Hong Kong

DA Data Augmentation

DAE Denoising Autoencoders

DBLP Digital Bibliography & Library Project

DL Deep Learning

DOC Deep Open Classification

DT Decision Trees

EM Expectation Maximization

FOPL First-order Predicate Logic

GA Genetic Algorithm

GDELT Global Database of Events, Language, and Tone

GTD Global Terrorism Database

IoT Internet of Things

ISIS Islamic State of Iraq and Syria

JJATT John Jay & ARTIS Transnational Terrorism Database

KNN K-Nearest Neighbors

LOF Local Outlier Factor

LR Logistic Regression

ML Machine Learning

MLE Maximum Likelihood Estimation

NASAA North American Securities Administrators Association

NB Naive Bayes

NLP Natural Language Processing

NN Neural Network

OC-SVM One-Class SVM

PSO Particle Swarm Optimization

RDF Resource Description Framework

SNA Social Network Analysis

SPC Statistical Process Control

START Study of Terrorism And Responses to Terrorism

SVM Support Vector Machine

TF-IDF Term Frequency-Inverse Document Frequency

TL Transfer Learning

TM Text Mining

UCSD University of California San Diego

WCAE Weighted Convolutional Autoencoder

WE Word Embedding

XSS Cross-Site Scripting

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction Générale

Sommaire

1.1 Contexte de la thèse

Le nombre de réseaux sociaux ne cesse d'augmenter régulièrement ! Twitter, Facebook, LinkedIn, Tumblr, Snapchat, Instagram, etc. Ce sont des plateformes idéales pour toucher à moindre frais un nombre très important d'internautes. Les réseaux sociaux sont une notion ambiguë, car au sens strict, un réseau social en ligne fait référence aux différentes relations que les gens entretiennent entre eux et à la manière dont elles sont structurées ; ces différentes relations aident à comprendre le comportement des individus. De nos jours, les gens se sont habitués à utiliser la notion de réseau social pour désigner une application dédiée à la communication ou plus particulièrement un service de réseau social qui, via Internet, permet de

maintenir la communication avec leurs familles, amis ou collègues, et de rencontrer de nouvelles personnes. Les réseaux sociaux permettent aux individus d'interagir avec d'autres individus inscrits sur le même réseau en échangeant des messages, en publiant des photos ou des vidéos publics ou privés. D'après une étude récente de Hootsuite et We Are Social sur l'usage du Web et des réseaux sociaux en 2020 [4], de la population mondiale qui comprend 7.75 milliards d'individus, 59% utilisent Internet et 49% sont actifs sur les réseaux sociaux. Comparé à l'année 2019, le nombre d'internautes a évolué de 7%, ce qui représente une augmentation importante sur une année, soit 321 millions d'individus supplémentaires. De ces réseaux sociaux, Facebook est classé en tête de liste avec 2.45 milliards d'utilisateurs, suivi de Youtube avec 2 milliards, puis WhatsApp avec 1.6 milliard d'utilisateurs. Ces chiffres continuent à croitre en 2021, surtout avec les restrictions liées à la pandémie mondiale COVID19. En effet, le pourcentage de la population utilisant les médias sociaux a atteint 56.8, par l'augmentation de 520 millions de nouveaux utilisateurs des plateformes sociales.

Les réseaux sociaux ne sont en fait que des systèmes complexes du monde réel, ils représentent des réseaux d'informations dynamiques hautement interconnectés [5][6] fournissant une disponibilité croissante de données [7]. L'analyse de ce type de réseaux nécessite une modélisation structurelle sous forme de graphes dont les nœuds représentent les entités dynamiques et les liens représentent les interactions entre ces entités. A titre d'exemple, pour Facebook, les utilisateurs peuvent être considérés comme étant les noeuds du graphe social et les relations entre ces utilisateurs comme étant les arêtes de ce dernier. Cette topologie de structure complexe peut être exploitée pour analyser, comprendre, prédire et optimiser le degré de câblage et le comportement de ces systèmes dynamiques [8]. Ainsi, les métriques graphiques fondamentales, telles que l'extraction des propriétés de chaque noeud (un ego), de son voisinage au premier niveau (un egonet),

de son voisinage au deuxième niveau (un super-egonet) et de la transitivité des arêtes, peuvent servir à collecter et interpréter des informations sur l'état ou le positionnement des différents composants du graphe [9]. Toutefois, cette complexité d'informations peut être mieux représentée à travers l'utilisation d'un graphe multidimensionnel réduisant le taux de manque de l'information. Ce type de graphe représentant l'interaction entre les constituants d'une variété de systèmes complexes [10], combinant plusieurs types de relations (dites couches ou dimensions), est particulièrement utile dans la modélisation des divers échanges de flux entre les composants d'un réseau. Par exemple, les auteurs enregistrés sur Digital Bibliography & Library Project (DBLP) peuvent entretenir plusieurs types de relations entre eux. En effet, deux auteurs peuvent être des co-auteurs, de même discipline, ayant publié un papier dans une même conférence, etc. Dans ce cas, DBLP serait le réseau, les auteurs seraient les noeuds et les différentes relations entre ces auteurs les dimensions.

Le concept des réseaux multidimensionnels a été largement étudié pour répondre à diverses problématiques. A titre d'exemple, dans [11], le réseau multidimensionnel a été utilisé pour modéliser les opérations ou les facteurs complexes des organisations dans le but d'élaborer de modèles d'évaluation du risque pays. Dans [12], afin de protéger et préserver les corridors patrimoniaux culturels, une structure multidimensionnelle a été développée pour connecter ces derniers en considérant différentes dimensions ou facteurs tels que le temps, l'espace, la culture ethnique, la culture religieuse et les différences d'altitude dans les parcelles de paysage. Une autre application de ces réseaux se manifeste dans la lutte contre les attaques prevenant de la forte connectivité des appareils de l'Internet des objets, où les fonctionnalités liant ces appareils présentent une abstraction de multiples couches figurant dans un réseau multidimensionnel [13]. D'autres ont adopté ce type de réseau pour la gestion d'une quantité importante de données

afin de faciliter la récupération de l'information de collaboration entre les chercheurs sur des réseaux multidimensionnels d'influence académique [14]. Les réseaux multidimensionnels ont été mis en oeuvre aussi dans la construction d'un réseau dynamique de relations biologiques et agronomiques multidimensionnelles afin de combiner des classifications des familles de gènes du sorgho (régulateurs/facteurs de transcription, enzymes actives sur les glucides, protéines kinases, etc.). Ce regroupement d'informations avait pour but d'extraire les principaux régulateurs qui influencent les voies métaboliques [15]. Dans le domaine de la médecine, des chercheurs ont incorporé une topologie multidimensionnelle pour mieux comprendre le degré d'interaction de quinze symptômes concomitants chez des patients en oncologie subissant une chimiothérapie [16]. Un autre travail oriente l'utilisation d'un réseau multidimensionnel de médias sociaux vers la prédiction des opinions communes entre les individus [17]. A notre connaissance, face à cette diversité et multiplicité de champs d'applications de réseaux complexes, peu de travaux dans la littérature se sont focalisés sur la détection d'anomalies. Les travaux existants traitent cette problématique principalement dans le contexte monodimensionnel.

La *détection d'anomalies* est l'identification d'éléments, d'événements ou d'observations rares qui soulèvent des suspicions compte tenu de leurs différences significatives par rapport à la majorité des autres données [18]. Généralement, les anomalies indiquent un problème tel qu'une fraude bancaire, un défaut structurel, un problème médical ou une erreur dans un texte. Les anomalies sont également appelées des valeurs aberrantes, du bruit, des écarts ou des exceptions. La détection d'anomalies est applicable dans divers domaines, tels que la détection d'intrusions [19], la détection de défauts, la surveillance de l'état du système, la détection d'événements dans des réseaux de capteurs, la détection de perturbations dans l'écosystème, ainsi que dans la détection des comportements anormaux dans des réseaux en ligne.

En fait, les réseaux en ligne exposent leurs utilisateurs, en particulier les jeunes car ils n'ont ni le recul ni l'expérience nécessaires pour discerner une situation à risque ou un contenu potentiellement préjudiciable et une multitude de dangers. Parmi ces dangers, nous citons le cyber-harcèlement ou le harcèlement sur Internet. Ce problème est apparu lors de l'avènement des réseaux sociaux, en tant que nouvel outil d'intimidation à travers la diffusion de textes malveillants, d'images ou de films diffamatoires. Il s'agit d'un acte agressif et intentionnel commis par un individu ou un groupe d'individus au moyen de communications électroniques. Un harcèlement peut se manifester par un message privé ou un statut public et peut représenter un risque lorsque l'auteur de ce message ou de ce statut est exposé à une large communauté. Plusieurs formes d'intimidation sont possibles : insultes, moqueries, menaces en ligne, propagation de rumeurs, images dégradantes ou réponse provocante à un message. Ces intimidations, dont le but est d'humilier une personne, mènent tout un processus de victimisation. Les réseaux en ligne, particulièrement les réseaux sociaux, sont aussi un espace d'usurpation d'identité. L'usurpation d'identité consiste à s'approprier délibérément l'identité virtuelle d'une autre personne, de pirater son nom d'utilisateur et son mot de passe pour utiliser son compte à des fins malveillantes. L'usurpateur peut également créer un nouveau compte à partir du compte de sa victime et le profiler avec des photos détournées.

En résumé, l'objectif majeur de cette thèse est de fournir des solutions qui aideraient à détecter les utilisateurs violents dans les réseaux sociaux en prenant en charge la nature dynamique et complexe de ces réseaux.

1.2 Motivations

Comme évoqué précédemment, bien que les réseaux en ligne soient un moyen de communication facile permettant un échange convivial, certains utilisateurs en profitent de manière néfaste. L'une des catégories les plus dangereuses d'utilisateurs est celle des groupes terroristes. Les plateformes de collaboration en ligne jouent évidemment un rôle clé dans la mondialisation du terrorisme. Ces plateformes s'avèrent être un redoutable outil de propagande et de recrutement. Le terrorisme, dans toutes ses manifestations, concerne tout le monde. L'utilisation d'Internet, à des fins terroristes, dépasse les frontières nationales amplifiant ses effets potentiels sur les victimes. Depuis la fin des années 1980 et avec l'apparition du Web, le développement de technologies, toujours plus sophistiquées, a permis la création d'un réseau de portée véritablement mondial, auquel l'accès est relativement aisé. La technologie Internet, qui facilite la communication, a également été détournée à des fins terroristes [20]. L'apparition de réseaux sociaux permet aux groupes terroristes d'interagir efficacement et souvent de manière anonyme, non seulement pour partager des documents et des informations, mais aussi pour établir une communauté d'individus connectés. D'après Evan Kohlmann [21], *"Aujourd'hui, 90% des activités terroristes sur Internet se déroulent à l'aide d'outils de réseautage social. Ces forums agissent comme un pare-feu virtuel pour aider à protéger l'identité de ceux qui y participent, et ils offrent aux abonnés la possibilité d'entrer en contact direct avec des représentants terroristes, de poser des questions, et même de contribuer et d'aider le cyberjihad."*

Le tableau 1.1 expose l'évolution considérable du nombre d'attentats terroristes et du nombre de morts sur la période 1983-2019. Les statistiques enregistrées montrent clairement la multiplication de ces chiffres dans une première phase (début des années 1980-2000), avec l'apparition du Web, puis dans une seconde phase (2001-2019), avec l'apparition des réseaux sociaux. En effet, le tableau 1.2 montre

une augmentation du nombre d'attaques de plus de 14 fois et une multiplication du nombre de décès de victimes par 23 après l'avénement des réseaux sociaux.

Table 1.1: Evolution des attentats terroristes dans le monde (1983-2019) avec l'apparition du Web et des réseaux sociaux.

Année	**Nombre d'attaques**	**Nombre de morts**
1983	24	365
1990	Emergence du Web	
1994	366	660
1997	142	1188
2001	203	3922
2004	Apparition de Facebook	
2004	287	2479
2005	Apparition de Youtube	
2006	Apparition de Twitter	
2007	563	3136
2010	Apparition de Instagram	
2011	870	3445
2012	2423	8279
2014	4835	28198
2016	2182	13866
2018	1606	8715
2019	826	4953

Table 1.2: Nombre total d'attentats terroristes et de morts dans le monde avant et après l'apparition des réseaux sociaux.

Période	**Nombre total d'attaques**	**Nombre total de morts**
1980-2000	2190	6818
2001-2019	31579	160278

Ces statistiques choquantes dressent un état des lieux dramatique ce qui incite à trouver des moyens automatisés permettant une lutte efficace contre ce flau. Ce combat commence par le développement de moyens efficaces de détection de

groupes terroristes. Un autre aspect s'avèrant important a motivé notre intérêt qui est la synchronisation des réseaux sociaux. Depuis 2012, Facebook et Instagram par exemple sont liés et chacun peut en quelques clics synchroniser ses profils sur les deux réseaux. Ainsi, à chaque fois qu'un utilisateur poste une photo sur sa page Instagram, le même contenu est automatiquement dupliqué sur son profil Facebook. Il en est de même, si un utilisateur synchronise son compte Twitter avec celui de LinkedIn. Compte tenu de cette synchronisation, l'utilisation d'un réseau multidimensionnel, pour représenter les comptes d'un utilisateur et les relations entre ces utilisateurs, présente plusieurs avantages via la minimisation des pertes d'informations, la réduction du temps d'analyse et la collecte de plus d'informations sur le même utilisateur.

1.3 Problématiques

Bien que la détection des comportements anormaux ait été bien étudiée dans les réseaux monodimensionnels [22][23][24][9][25][26], peu d'approches ont été élaboré à ce sujet dans les réseaux multidimensionnels [27]. Particulièrement, plusieurs approches et méthodes ont été développées dans une optique de détection des comportements terroristes telles que la surveillance manuelle et les pare-feu. Certains de ces efforts se sont concentrés sur l'utilisation d'approches basées sur l'analyse de l'activité des utilisateurs [28][29], tandis que d'autres se sont focalisés sur l'analyse de la structure topologique des réseaux sociaux réunissant les différentes communautés d'utilisateurs [30][31]. Dans ce contexte, les méthodes qui utilisent des métriques graphiques (méthodes structurelles) sont les plus efficaces pour l'analyse de relations complexes comme c'est le cas d'un réseau multidimensionnel [32][33][34]. Bien que les travaux de recherche cités ci-dessus aient contribué à résoudre, en partie, les problèmes de détection des comportements anormaux,

des défis en rapport avec la dimension du graphe des réseaux sociaux, la multimodalité des données analysées et la structure communautaire dynamique, demeurent encore non résolus, et notamment, parce-que :

— *La structure des réseaux sociaux est monodimensionnelle :* Seule la contribution [27] s'est focalisée sur la détection des anomalies dans des réseaux multidimensionnels, ce qui constitue un manque dans la littérature. L'aspect multidimensionnel des réseaux doit être pris en considération dans le but d'améliorer la précision et la performance de l'analyse. L'exploration des réseaux multidimensionnels constitue un avantage majeur pour la récupération d'une quantité importante d'informations sur une même instance existante sur le réseau. De plus, tenir compte uniquement de la structure topologique multidimensionnelle sans avoir recours à l'aspect communautaire est insuffisant. En effet, la détection des différentes communautés dans les réseaux sociaux permet de rendre plus clair le type de relations entre les utilisateurs. À notre connaissance, aucun travail de la littérature n'a abordé ces deux aspects.

— *Les données analysées sont homogènes :* Peu de contributions [35][28][36][37] se sont intéressées à la manipulation et l'extraction des données multimodales. Les diverses catégories de données présentent sur les réseaux sociaux sont hétérogènes et complémentaires. Un tel traitement approfondi sur la structure des réseaux sociaux nécessite obligatoirement d'être capable d'exploiter tous les types de données afin d'enrichir et garantir des résultats d'analyse précis et de haute performance.

— *Les méthodes sont comportementales ou structurelles :* Les contributions proposées dans la littérature traitent généralement, soit l'aspect comportemental des individus sur le réseau, soit l'aspect structurel d'interrelation entre des individus. Ces deux aspects sont complémentaires et doivent être

combinés pour avoir une vue globale sur les activités et les relations d'un individu à la fois.

— *Le comportement est statique :* La plupart des contributions développées à ce propos ne se focalisent pas sur le comportement dynamique des utilisateurs dans les réseaux sociaux. La complexité de l'analyse de tels comportements est considérée l'obstacle majeur. Il est primordial de tenir compte de la nature dynamique et variable dans le temps des profils pour traiter les nouveaux utilisateurs et leurs relations.

Vu les limites révélées des approches de détection des comportements anormaux déjà proposées, **l'idée principale de cette thèse est de mettre l'accent sur l'élaboration d'une nouvelle solution communautaire, dynamique, combinant les propriétés structurelles complexes et comportementales pour analyser et traiter plusieurs types de données.**

1.4 Objectifs et contributions

Les utilisateurs des réseaux sociaux établissent des liens d'*amitié virtuelle* ou suivent les activités des uns et des autres afin de faire connaissance ou de trouver des centres d'intérêts communs. Ces intérêts communs peuvent être de nature illicite menant jusqu'à des groupes de cybercriminalité et de terrorisme. Ces groupes sont généralement animés par des personnes influentes sachant coopter d'autres personnes présentant des prédispositions particulières. [38][39][40]. À la lumière de ce qui précède, les principaux objectifs de cette thèse sont énumérés comme suit :

— Fournir un modèle générique capable d'analyser de données multimodales dans une structure multidimensionnelle

— Rationaliser et améliorer les moyens d'examen et d'analyse des interactions et de la détection des communautés sur un réseau multidimensionnel

— Prendre en charge l'importance de la combinaison de l'aspect structurel communautaire et l'aspect comportemental analysant divers types de données sur une structure topologique multidimensionnelle dynamique

Ce travail de thèse répond aux objectifs ci-dessus en proposant quatre contributions majeures.

1. La première contribution propose une étude de la littérature qui passe en revue diverses méthodes d'exploration de données dans le contexte de la détection des comportements anormaux dans les réseaux sociaux. Cette étude fournit une meilleure évaluation pouvant faciliter la compréhension de ce domaine. Nous avons discuté trois types de méthodes, à savoir : les méthodes comportementales, les méthodes structurelles et les méthodes hybrides, où nous nous sommes concentré sur leurs forces et leurs faiblesses en fournissant un ensemble de critères que nous considérions comme des mesures de performance. Ces mesures se résument en quatre points, à savoir : la capacité de l'analyse d'une structure complexe comme celle des réseaux sociaux, la prise en charge de divers types de données en analyse, la considération de la nature dynamique des réseaux et l'adaptabilité avec une structure communautaire multidimensionnelle. Malgré les nombreux efforts consentis sur ce sujet, nous avons conclu qu'il n'existe pas de méthodes efficaces, qui répondent pleinement aux différentes exigences. Avec l'avènement de nouvelles techniques d'apprentissage et de détection d'anomalies, nous avons constaté que cet axe de recherche est toujours actif où de futurs travaux doivent nécessairement être mis en œuvre pour étendre les méthodes hybrides, afin qu'elles soient capables de s'affranchir de la structure multidimensionnelle des réseaux sociaux et d'ajouter des contraintes temporelles pour garantir le dynamisme.

2. La deuxième contribution présente un modèle de détection et de prédiction des comportements violents sur une structure sociale monodimensionnelle (Twitter). Ce choix de type de structure a été motivé par le fait que nous puissions montrer l'importance de l'analyse d'une structure multidimensionnelle dans nos prochaines contributions. Ce modèle propose une nouvelle méthode pour prédire les utilisateurs suceptibles être terroristes. Cette méthode prend en charge l'extraction et l'analyse de deux types de données seulement, textuelles et images, afin de prouver de même dans nos prochaines contributions que la considération d'un type supplémentaire de données conduit à de meilleurs résultats. Par conséquent, pour valider la bonne performance de notre méthode, nous avons entrainé le modèle construit avec des données réelles sur le terrorisme extraites de plusieurs sources de données (en ligne et hors ligne). Par la suite, nous avons testé ce dernier sur un ensemble de données réelles provenant de Twitter. À des fins de visualisation, nous avons construit un graphe social qui rend les données de sortie plus facilement compréhensibles et interprétables par un humain. Ce graphe s'avére utile pour toute autre analyse d'un réseau social. Les résultats obtenus ne sont pas très optimaux, ils n'ont pas dépassé une précision de 72%.
3. La troisième contribution propose une méthode de détection des comportements atypiques sur une structure topologique multidimensionnelle (traitant la synchronisation des trois réseaux sociaux, à savoir : Facebook, Twitter et Instagram). Cette méthode s'intéresse à l'exploration des relations entre les différents composants d'un réseau afin d'extraire et calculer des métriques graphiques. Ces métriques sont définies pour identifier les différentes communautés sur le réseau afin d'estimer un score d'anomalie à chaque noeud suite à son degré d'implication dans sa communauté. Les

scores d'anomalies sont calculés par l'application d'une nouvelle fonction que nous avons proposée. Nous avons eu recours, suite à cette estimation de scores, à l'identification automatique des noeuds atypiques en utilisant le modèle probabiliste de la distribution Bêta. Pour valider notre travail, nous avons testé cette méthode sur un graphe social multidimensionnel où nous avons généré une matrice d'adjacence qui montre clairement les scores d'anomalies triés par ordre croissant. Les scores les plus faibles sont classés en haut de la matrice. Par conséquent, ils sont considérés comme étant des anomalies connectés de manière clairsemée au réseau, alors que les nœuds normaux sont étroitement connectés et se manifestent dans la matrice par des régions denses. Bien que les résultats obtenus soient très motivants, l'aspect comportemental est manquant. Nous avons combiné dans notre quatrième contribution les deux aspects comportemental et structurel multidimensionnel afin d'aboutir à de nouveaux résultats plus optimaux.

4. La quatrième contribution consiste en l'élaboration d'un framework d'identification des personnes dont le comportement laisse à penser qu'il puisse s'agir d'activités potentiellement dangereuses. Ce framework analyse les noeuds atypiques du réseau pour s'assurer du caractère violent ou non de leur comportement. Pour aboutir à cet objectif, nous avons construit un modèle multidimensionnel qui s'exécute sur des données traitant d'un nouveau type par rapport à la deuxième contribution, à savoir, les informations générales. Ce modèle permet la sélection des caractéristiques publiques des personnes et la transmission de ces caractéristiques à trois sous-modèles. Chaque sous-modèle est réservé à l'analyse d'un type particulier de données. Un score est calculé pour chaque comportement analysé en fonction d'un poids déterminé et attribué automatiquement à chaque sous-modèle, pour être enfin classé par un seuil de décision calculé automatiquement

suite à la précision des algorithmes d'apprentissage. Les expérimentations de ce travail ont montré de bons résultats (95% de précision) provenant du test de notre framework sur un ensemble de données réelles extraites par des moyens différents des mêmes trois réseaux sociaux populaires que notre troisième contribution.

1.5 Publications

— Journal Articles

1. **Nour El Houda Ben Chaabene**, Amel Bouzeghoub, Ramzi Guetari, Henda Hajjami Ben Ghezala : New Deep Learning Framework for Detecting the Behavior of a Terrorist Group on a Multidimensional Network Using Multimodal Data. Expert Systems With Applications, 2021.
2. **Nour El Houda Ben Chaabene**, Amel Bouzeghoub, Ramzi Guetari, Henda Hajjami Ben Ghezala : Deep learning methods for anomalies detection in social networks using multidimensional networks and multimodal data : a survey. Multimedia Systems, 2021.

— Conference Proceedings

1. **Nour El Houda Ben Chaabene**, Amel Bouzeghoub, Ramzi Guetari, Henda Hajjami Ben Ghezala : Applying Machine Learning Models for Detecting and Predicting Militant Terrorists Behaviour in Twitter. In IEEE International Conference on Systems, Man, and Cybernetics (SMC-2021), Melbourne, Australia, October 17-20, 2021.
2. **Nour El Houda Ben Chaabene**, Amel Bouzeghoub, Ramzi Guetari, Samar Balti, Henda Hajjami Ben Ghezala : Detection of Users' Abnormal Behavior on Social Networks. In 34th International Conference on

Advanced Information Networking and Applications (AINA-2020), Advanced Information Networking and Applications, vol 1151. pp. 1-13, Caserta, Italy, April 15-17, 2020.

1.6 Organisation de la thèse

Cette thèse de doctorat se divise en huit chapitres et est structurée comme suit :

Chapitre 1 : Introduction Générale. Ce chapitre présente le contexte général dans lequel s'inscrit notre travail, à savoir la détection des comportements violents dans les réseaux sociaux. Ce chapitre traite la motivation générale de notre travail et met en évidence les défis de l'analyse des réseaux complexes. Tout d'abord, il se concentre sur un aperçu des travaux limités de la littérature et des problèmes abordés dans cette thèse. Ensuite, il précise les principaux objectifs et contributions de ce travail doctoral.

Chapitre 2 : Qu'est-ce qu'un comportement anormal ?. Ce chapitre décrit les concepts de base de notre recherche. Tout d'abord, il définit les notions de modélisation du comportement d'un individu dans les réseaux sociaux. Ensuite, il aborde les différentes définitions de la littérature d'un comportement anormal des individus en décrivant leurs typologies. Enfin, il présente les catégories des menaces en mentionnant leurs variétés.

Chapitre 3 : Techniques d'analyse des réseaux sociaux. Ce chapitre traite les différentes techniques exploitées pour l'analyse de la structure des réseaux sociaux. Ces techniques sont triées par le type de données qu'elles utilisent.

Chapitre 4 : Méthodes de détection d'anomalies dans les réseaux sociaux : Etat de l'art. Ce chapitre catégorise les méthodes de détection des

anomalies et identifie leurs limites en s'appuyant sur un ensemble de critères d'évaluation. Cette étude de la littérature a permis de mettre en évidence les limites des travaux existants et de positionner les contributions de la thèse.

Chapitre 5 : Modèle de détection et de prédiction des comportements anormaux sur Twitter. Ce chapitre propose une première contribution à notre problèmatique. Cette solution met en oeuvre un nouveau modèle qui se concentre sur l'analyse de la structure monodimensionnelle de Twitter afin de détecter et prédire les comportements terroristes.

Chapitre 6 : Méthode de détection des comportements anormaux sur la base de l'analyse des relations dans une structure multidimensionnelle. Ce chapitre présente une deuxième contribution à la problèmatique de cette thèse. Il s'agit d'une nouvelle méthode exploitant les relations entre les utilisateurs dans un réseau multidimensionnel pour déduire le type de comportement analysé.

Chapitre 7 : Framework hybride de détection des comportements anormaux sur un réseau multidimensionnel utilisant des données multimodales. Ce chapitre développe la troisième contribution de la thèse dont l'objectif est d'intégrer et d'améliorer les résultats des deux contributions précédentes en prenant en compte à la fois l'aspect structurel et comportemental et en explorant les relations entre individus par le biais d'une structure multidimensionnelle afin d'analyser le comportement de chaque individu sur la base de ses activités.

Chapitre 8 : Conclusion et travaux futurs. Ce chapitre fournit un résumé des résultats scientifiques apportés par cette thèse et propose des directions de recherche en rapport avec des problèmes connexes.

Chapitre 2

Qu'est-ce qu'un comportement anormal ?

Sommaire

2.1 Introduction

La cybercriminalité est un terme qui couvre un large éventail d'activités criminelles au moyen d'un ordinateur. Elle désigne l'acte délictueux utilisant le cyberespace comme moyen de communication. De nos jours, le monde connaît une croissance exponentielle dans le cyberespace et une ascension importante des activités sur Internet. Il va sans dire qu'une telle croissance phénoménale de l'accès à l'information conduit, d'une part à donner plus de pouvoir aux individus et aux organisations, et d'autre part à de nouveaux défis pour les gouvernements et les citoyens. Le Web, et plus particulièrement les réseaux sociaux, ont créé un monde virtuel où des communautés peuvent se former comme dans le monde réel. Ces communautés sont formées par des individus physiques qui peuvent avoir des comportements totalement hétérogènes.

2.2 Qu'est ce que le comportement d'un individu ?

Quelle définition retenir de l'expression «comportement d'un individu» de la littérature ? De manière générale, le comportement d'un individu se caractérise par l'ensemble des réactions adoptées par une personne, dans son environnement et en réponse à des situations données. Albu [41] et Zhang et al. [42] ont indiqué que le comportement humain est défini par des caractéristiques corporelles spécifiques et des mouvements tels que des gestes et des plaisirs visuels du visage qui

aident à comprendre les relations sociales entre les individus. Banovic [43], de son côté, a défini le comportement d'un individu comme une séquence d'actions qui changent d'un état à un autre jusqu'à atteindre un état but. Le comportement d'un être humain s'adapte généralement au rythme de la vie et à l'évolution socio-culturelle. Dans notre société moderne, nous pouvons citer un certain nombre de comportements liés à divers champs et pratiques.

- **Le comportement routinier :** il définit presque tous les aspects de la vie des gens. Les routines sont un type de comportement délibéré composé d'actions dirigées vers un but que les gens acquièrent, apprennent et développent à travers des activités répétées [43].
- **Le comportement de santé :** un comportement de santé est tout comportement ou activité qui fait partie de la vie quotidienne, mais qui affecte l'état de santé de la personne. Presque tout comportement ou activité peut avoir une influence sur la santé, et dans ce contexte, il peut être utile de considérer les comportements liés à la santé comme faisant partie intégrante du mode de vie d'un individu ou d'un groupe.
- **Le comportement du consommateur :** Le comportement du consommateur désigne les réactions d'un individu considéré comme client réel ou potentiel d'une entreprise en fonction de stimulis, par exemple lors de sa visite à un magasin. La stimulation de la demande du consommateur dépasse l'idée naïve de satisfaction des besoins au sens strict. L'analyse du comportement du consommateur cherche à identifier les déterminants de ce comportement (besoins, motivations, attentes, critères de choix, etc.) en vue de permettre à l'entreprise de s'y adapter ou de les influencer dans une vision concurrentielle [44].
- **Le comportement politique :** c'est l'ensemble des pratiques sociales et

des actes liés à la vie politique. C'est la participation des individus aux élections et, plus généralement la participation à des manifestations ou mouvements sociaux, l'adhésion à un parti politique, etc. Les individus, par ce comportement, tentent d'influencer les gouvernants.

Le comportement des individus est souvent lié aux termes confiance, réputation, influence et popularité dans le cadre de l'analyse des réseaux sociaux. Les auteurs [45] soulignent les caractéristiques les plus importantes que les définitions de ces notions devraient inclure et font l'analyse comparative des termes les plus souvent confondus (confiance vs réputation, et influence vs popularité). Une classification globale, concernant leur niveau abstrait, définit les termes et les distingue les uns des autres. Pour récapituler, la définition du comportement d'un individu est l'ensemble des réactions, des activités ou des opérations dépendant d'un contexte bien défini. Par exemple, le comportement d'un utilisateur de réseaux sociaux apparaît à partir de ses messages, commentaires, partages, etc. Par ailleurs, cette définition rejoint les deux définitions psychologiques et sociologiques du terme "comportement". En psychologie, le comportement est un ensemble de mouvements physiques et de réactions motrices. Il est défini en termes de propriétés psychochimiques qui reflètent que sa modélisation est difficile, voire impossible. En sociologie, le comportement est défini par l'ensemble des réactions adaptatives observables dans le temps. Du point de vue informatique, la modélisation du comportement peut être formalisée aisément par un modèle mathématique.

2.3 Modèles informatiques représentants les comportements des individus dans les réseaux sociaux

Le comportement d'un utilisateur de médias sociaux représente la manière dont les personnes interagissent entre elles et se charge de l'énumération des diverses activités sociales faites par ces personnes. Pour mieux comprendre le comportement d'un individu, il est nécessaire d'en savoir davantage sur les modèles informatiques capables de modéliser leur comportement.

2.3.1 Représentation vectorielle

La représentation la plus facile à adapter pour la modèlisation d'un comportement est la représentation vectorielle. Elle permet de définir l'ensemble des caractéristiques d'un profil utilisateur et d'appliquer un poids à chacune de ces caractéristiques [46]. Ces poids révèlent l'importance d'une caractéristique par rapport aux autres. Suite à cela, nous pouvons considérer que la description d'un profil est basée sur un ensemble de mots-clés que nous pouvons pondérer [47].

2.3.2 Représentation hiérarchique

La représentation d'un comportement par une structure hiérarchique permet de définir un ensemble de concepts qui assurent la généralité du profil de l'utilisateur. Par exemple, si un utilisateur est intéressé par le sport, il est probablement également intéressé par un type spécifique de sport comme le tennis. L'identification de ces concepts hiérarchiques peut avoir deux formes ; une identification statique où les concepts sont fixés dès le départ et ne peuvent pas être modifiés [48] et une identification dynamique où les concepts sont rectifiés à chaque fois qu'il y a un

changement dans le comportement de l'utilisateur [49]. La représentation hiérarchique est souvent employée pour modéliser les centres d'intérêt des personnes.

2.3.3 Représentation ontologique

L'utilisation d'un moyen de formalisation comme les ontologies permet d'obtenir une description enrichie du profil de l'utilisateur, tout en permettant d'effectuer des raisonnements. D'après Sakthi et Revathy [50], en tant que modèle de description et de formalisation des connaissances, les ontologies sont largement utilisées pour représenter le comportement de l'utilisateur en associant des liens sémantiques entre les concepts détectés [48][51][52].

2.3.4 Représentation graphique

Le graphe est un outil mathématique efficace pour la représentation des relations entre les utilisateurs dans les réseaux sociaux où les nœuds représentent les utilisateurs et les arcs représentent les relations entre ces utilisateurs [53]. Deux types de graphe peuvent être mis en jeu [53] :

— Les graphes non orientés comme par exemple les graphes d'amitié et les graphes d'interaction : Un graphe d'amitié désigne par ses noeuds les utilisateurs et par ses arcs les relations amicales entre les utilisateurs. Par contre, un graphe d'interaction reflète seulement les interactions réelles des utilisateurs dans les réseaux sociaux, où seule une interaction visible entre deux utilisateurs (par exemple, laisser un commentaire à propos d'une photo) peut créer un arc dans le graphe.

— Les graphes orientés comme les graphes latents : pour cette instance de graphe orienté, un arc dirigé du noeud A vers un noeud B désigne que l'utilisateur A à consulté le profil de l'utilisateur B. Par conséquent, le nombre d'arcs sortants d'un nœud représente le nombre de visites effectuées

par cet utilisateur aux autres utilisateurs et le nombre d'arcs entrants dans ce noeud représente le nombre de visiteurs de ce profil.

2.3.5 Clickstream

Le clickstream trace une représentation des chemins empruntés par les utilisateurs sur les site Internet. Et puisque le comportement de l'utilisateur est défini en fonction de ses centres d'intérêts, les pages qu'il a visité et ses publications [54], nous pouvons utiliser le modèle Clickstream comme moyen de modélisation du comportement des utilisateurs des réseaux sociaux. Généralement, les flux de clics d'un utilisateur dans un site Web incluent une identification du nombre de visites, d'un identifiant unique pour chaque visite, de la date et l'heure de la visite, de l'adresse IP de l'appareil utilisée lors de la visite, etc. Par exemple, les entreprises de commerce électronique utilisent souvent les flux de clics pour avoir un avis sur les préférences des clients. Le Clickstream a comme objectif principal de collecter des informations supplémentaires et précises sur l'internaute afin de lui fournir des suggestions personnalisées [55].

2.4 Qu'est ce qu'un comportement anormal ?

Selon l'Organisation mondiale de la santé, et en ce qui concerne le comportement anormal d'un être humain : "Les troubles du comportement se caractérisent par un changement de pensée, d'humeur ou de comportement qui ne correspond plus aux normes ou croyances culturelles." Par conséquent, le comportement anormal est invariablement bizarre, il peut être classé comme un comportement irrégulier et souvent illégal. A titre d'exemple, nous pouvons citer la fraude à la carte de crédit, les intrusions informatiques, les activités terroristes, etc. Dans la littérature, les auteurs ont fourni plusieurs définitions du comportement anormal des

utilisateurs des réseaux sociaux. Selon Chouchane et Bouguessa [27], un comportement anormal est défini par un écart par rapport au comportement normal connu à l'avance. Par conséquent, un comportement anormal peut être considéré comme étant une anomalie, un état anormal, un comportement inattendu ou une valeur aberrante. Une autre définition de Kaur et Singh [1] considère que le comportement anormal résulte d'activités inhabituelles, illégales ou erratiques, qui différent diamétralement des activités associées à des comportements qualifiés de normaux selon les pratiques en vigueur dans un cadre socioculturel déterminé. À son tour, Grubbs [18] a défini un comportement anormal par une valeur aberrante qui semble s'écarter de manière significative des autres membres de l'échantillon dans lequel il se produit. Selon Barnett et Lewis [56], l'état anormal d'un individu est une observation ou un sous-ensemble d'observations qui semblent être incompatibles avec le reste de cet ensemble de données. Aggarwal et Yu [57] considèrent les comportements anormaux comme des points de bruit, qui se trouvent en dehors d'un ensemble défini de clusters, ou comme des points, qui sont en dehors de l'ensemble de clusters mais sont également séparés du bruit. Quant à Chandola et al. [58], les anomalies sont définies comme des modèles de données qui ne se conforment pas à une notion bien définie de comportement normal. Savage et al. [59] ont noté que les anomalies sont des régions du réseau dont la structure diffère de celle prédite dans le modèle normal.

2.5 Typologies d'un comportement anormal

D'après les travaux abordés dans la littérature, nous distinguons deux différentes catégories d'anomalies [58][1]. La première catégorie met en question les natures d'anomalies et la deuxième catégorie met l'accent sur les types d'anomalies qui se basent sur le type de données disponibles dans la structure des réseaux.

2.5.1 Natures d'anomalies

Nous pouvons répartir les anomalies en quatre classes [1] : les anomalies ponctuelles, contextuelles, collectives et horizontales. Il est important d'identifier la nature d'une anomalie afin de bien pouvoir choisir la méthode la plus adaptée à sa détection. La nature d'anomalie considérée dépend du problème en question.

— **Anomalie ponctuelle :** Si une instance de données est individuelle, elle est considérée comme anormale par rapport au reste des données. De ce fait, elle est appelée une anomalie ponctuelle ou une anomalie globale. Cette nature d'anomalies est la plus simple à identifier, il suffit de trouver les moyens pour détecter la déviation de l'instance anormale par rapport aux autres. A titre d'exemple, si nous considérons qu'un noeud d'un réseau ne peut être classé "normal" que s'il est connecté au minimum à trois voisins. La figure 2.1 montre que l'ensemble $E1$ présente des noeuds à comportement normal par rapport à l'ensemble $E2$ qui contient des noeuds anormaux vu leur séparation dans l'espace.

— **Anomalie contextuelle :** Si une instance de données est considérée comme anormale dans un contexte spécifique, mais pas autrement, alors il s'agit d'une anomalie contextuelle (également appelée une anomalie conditionnelle). A titre d'exemple, la température peut être considérée comme une anomalie contextuelle, dans le cadre de si la température est de 30 ° C en hiver à Toronto. Chaque instance de données est définie à l'aide des deux ensembles d'attributs :

  — Les attributs contextuels : se sont les attributs qui définissent le contexte de l'objet. Dans l'exemple précédent, les attributs contextuels sont la saison et le lieu.

  — Les attributs de comportement : se sont les attributs qui définissent les caractéristiques d'un objet et qui en quelque sorte, sont consacrés à

l'identification du comportement anormal d'un objet d'après son contexte. Pour le même exemple cité précédemment, la température et l'humidité peuvent être considérées comme des attributs de comportement.

— **Anomalie collective :** Si un groupe de données est anormal par rapport au reste des données, les données de ce groupe sont définies comme des anomalies collectives. La figure 2.2 représente un ensemble A de noeuds considérés comme des anomalies collectives car ils ont une densité très élevée par rapport à celle des autres noeuds. Il faut bien savoir qu'un noeud individuel dans le groupe A n'est pas une anomalie par rapport aux noeuds de son groupe. Considérons l'exemple illustré dans la figure 2.3 d'une série temporelle contenant une sous-série anormale parce qu'elle est différente par rapport à l'ensemble de sous-séquences de la série temporelle. Ceci peut correspondre que l'état du compteur est en arrêt, qui échoue à remonter des données.

— **Anomalie horizontale :** Les anomalies horizontales [60] existent principalement dans les réseaux sociaux qui décrivent la présence d'anomalies en fonction de sources hétérogènes de données disponibles. Par exemple, un même utilisateur peut être présent dans différentes communautés sur différents réseaux sociaux. De même, un utilisateur peut avoir des amis similaires sur un certain nombre de réseaux sociaux (par exemple Facebook, Instagram) mais des amis complètement différents dans un autre réseau social (par exemple Twitter). Cela décrit une activité inhabituelle qui peut être considérée comme anormale.

Après l'étude des trois natures d'anomalies, nous avons interprété que pour répondre à notre problématique générale de la thèse, nous devrions mettre la lumière sur la détection d'anomalies contextuelles et collectives. Ce choix s'explique ainsi :

— D'une part, l'identification des comportements anormaux à travers l'analyse

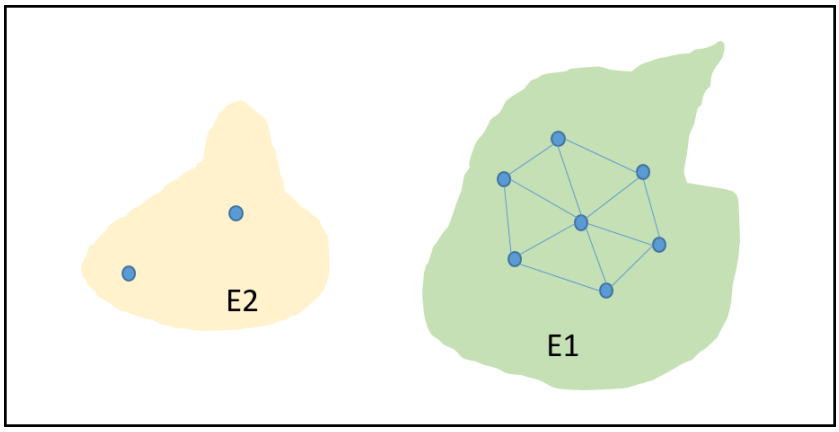

Figure 2.1: Anomalie ponctuelle.[source : [1]]

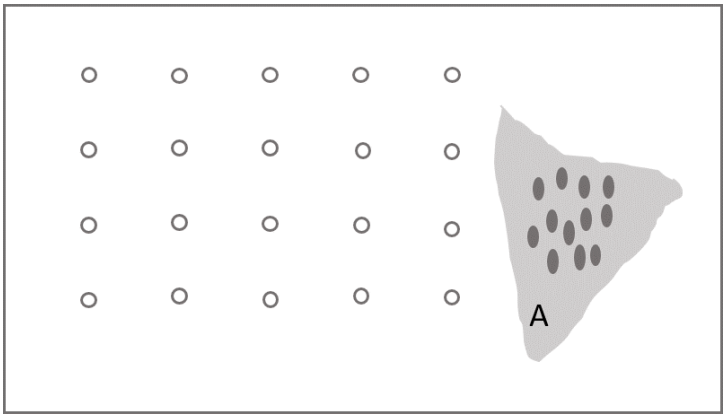

Figure 2.2: Anomalie collective.[source : [1]]

des relations entre les utilisateurs exige la construction de communautés d'utilisateurs. Or par définition, une communauté regroupe les personnes qui partagent presque les mêmes comportements. Ainsi, si une communauté est anormale, elle va regrouper un ensemble d'anomalies collectives.

— D'autre part, si nous nous intéressons à la détection des comportements terroristes, il est primordial de spécifier le contexte de ces comportements car les attributs contextuels et comportementaux qui définissent ce type spécifique de comportement sont différents par rapport à ceux d'un autre comportement spécifique.

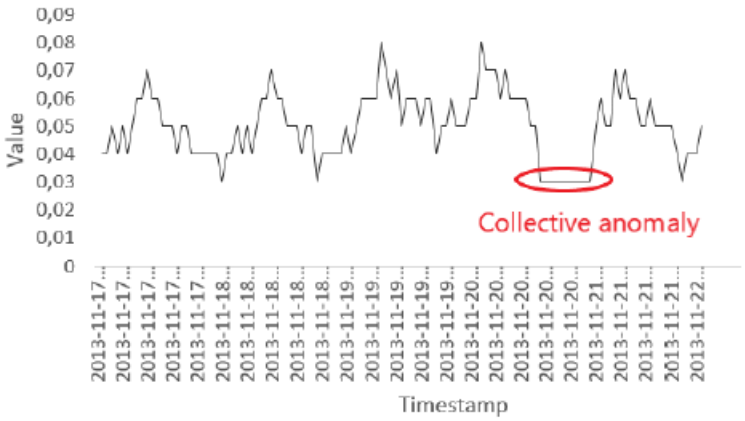

Figure 2.3: Anomalie collective correspondant à l'arrêt d'un compteur.[source : [2]]

2.5.2 Types d'anomalies

Dans le contexte de détection d'intrusions sur un réseau social, le type des données collectées permet de mettre l'accent sur le type de techniques abordées dans le processus de détection. Il existe deux types de techniques de détection d'anomalies : les techniques supervisées et les techniques non supervisées pour la détection d'anomalies étiquetées et non étiquetées respectivement [1].

— **Anomalies étiquetées :** Les techniques supervisées de détection d'anomalies nécessitent un ensemble de données où ces données doivent être étiquetées par "normales" ou "anormales" et impliquent l'entrainement d'un classificateur.

— **Anomalies non étiquetées :** La détection des anomalies non étiquetées nécessite des techniques non supervisées. Le principe de ces techniques est que la plupart des instances des données sont supposées normales et que le processus de recherche d'anomalies se concentre sur les instances qui ne ressemblent pas au reste des données. Les techniques non supervisées de détection d'anomalies sont principalement utilisées en matière de clustering et destinées à regrouper un ensemble de données hétérogènes sous forme de

sous groupes homogènes ou liés par des caractéristiques communes.

Les données disponibles réelles, que nous pouvons collecter de la structure des réseaux sociaux, sont pour la plupart des données non étiquetées. De ce fait, nous nous sommes intéressés par l'utilisation des techniques supervisées et non supervisées dans le cadre de la résolution de la problématique de cette thèse.

2.6 Menaces dans les réseaux sociaux

Dans un monde de plus en plus marqué par l'augmentation phénoménale de l'usage des réseaux sociaux, plusieurs utilisateurs se trouvent involontairement exposés à diverses menaces qui peuvent affecter leur sécurité et leur vie privée. Nous pouvons distinguer quatre categories de menances [3][61] : les menaces classiques, les menaces modernes, les menaces combinées et les menances ciblant les enfants. La figure 2.4 illustre les menaces mentionnées ci-dessuset leurs variétés.

2.6.1 Menaces classiques

Les menaces classiques sont devenues de plus en plus virales par leur forte propagation parmi les utilisateurs des réseaux sociaux. En effet, par des outils et techniques comme les logiciels malveillants, le phishing, le spam, le scripting croisé, un attaquant peut profiter non seulement des informations personnelles publiées d'un utilisateur dans un réseau social, mais aussi de celles de ses amis, simplement en adaptant la menace aux informations personnelles de ce dernier. Généralement, cette catégorie de menaces cible les informations essentielles et quotidiennes des internautes, telles que les numéros de carte de crédit, les mots de passe des comptes, etc. Elle peut également concerner les informations d'identification volées de l'utilisateur victime pour publier des messages en son nom ou même modifier

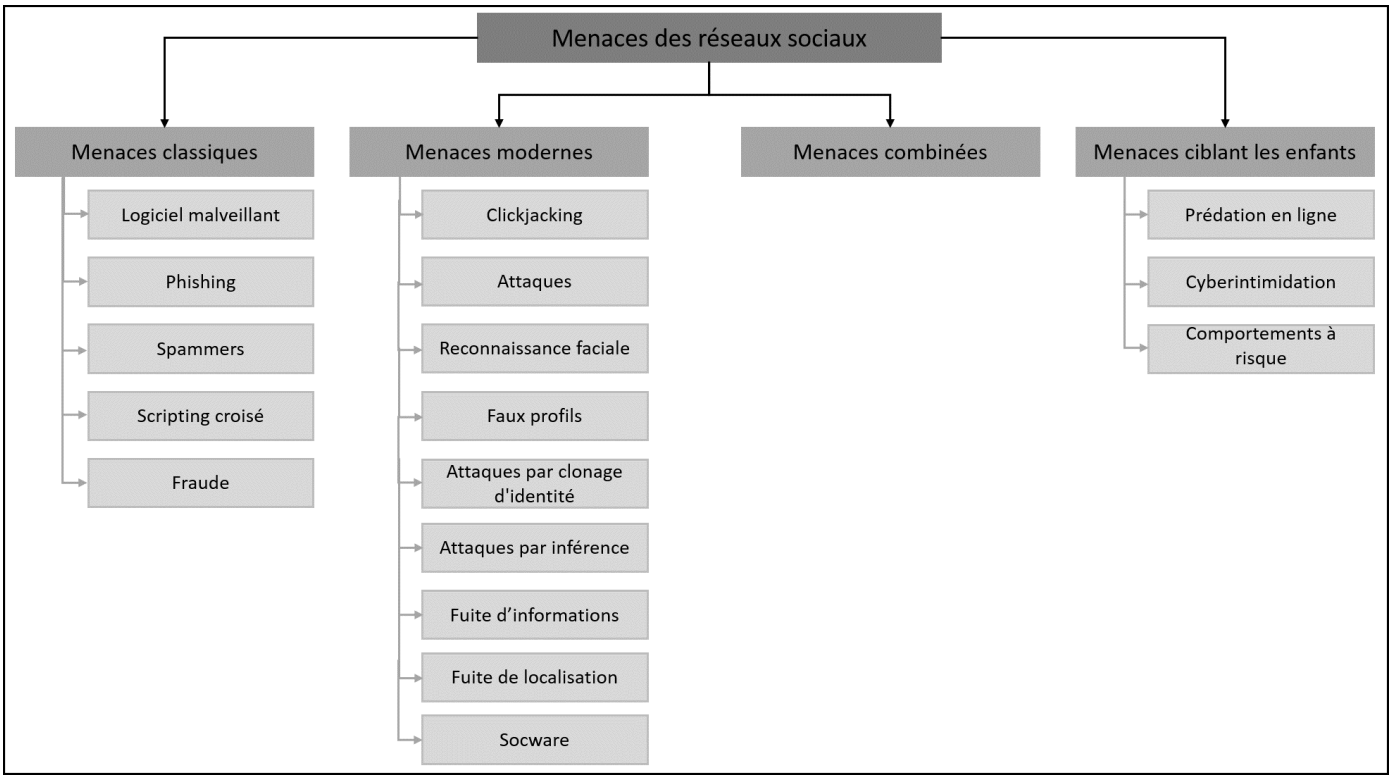

Figure 2.4: Menaces de réseaux sociaux en ligne et leurs variétés.[source : [3]]

ses informations personnelles. Dans ce qui suit, nous décrivons les principales varités de menaces classiques et comment ce type de menaces peut réellement nuire la sucurité et la vie privée de ses victimes.

Logiciels malveillants. Un malware est un logiciel malveillant développé dans le but de collecter les informations d'identification d'un utilisateur en ligne sur son ordinateur et d'accéder à ses informations personnelles. Les logiciels malveillants dans les réseaux sociaux utilisent la structure de ces derniers pour se propager parmi les utilisateurs et leurs amis connectés. Un logiciel malveillant peut même utiliser les informations d'identification collectées pour se faire passer pour l'utilisateur attaqué et envoyer des messages contagieux à ses amis en ligne.

Phishing. Les attaques de phishing sont une forme d'ingénierie sociale visant à acquérir des informations sensibles et privées des utilisateurs en se faisant passer pour un tiers de confiance. Une étude [62] a montré que les utilisateurs qui interagissent sur les sites de réseaux sociaux sont plus susceptibles de tomber dans le piège du phishing en raison de leur nature sociale et de leur confiance.

Spammeurs. Les spammeurs sont des utilisateurs qui se servent des systèmes de messagerie électronique pour envoyer des messages indésirables à d'autres utilisateurs. Les réseaux sociaux en ligne peuvent être utilisés par des spammeurs pour envoyer des messages publicitaires à d'autres utilisateurs en créant de faux profils [22], ou ajouter des messages de commentaires aux pages qui sont consultées par de nombreux utilisateurs du réseau.

Cross-Site Scripting (XSS). Une attaque XSS est une attaque contre les applications Web. L'attaquant qui utilise le XSS exploite la confiance du client Web dans l'application et fait en sorte que le client exécute un code malveillant capable de collecter des informations sensibles. Les réseaux sociaux en ligne, qui sont des types d'applications, peuvent facilement être victimes d'attaques XSS. En outre, les attaquants peuvent utiliser une vulnérabilité XSS combinée à l'infrastructure de ces réseaux pour créer un ver XSS qui peut se propager de manière virale parmi les utilisateurs du réseau social [63].

Fraude sur Internet. La fraude sur Internet, également appelée cyberfraude, consiste à utiliser l'accès à Internet pour escroquer les gens ou en tirer profit. Avant, les escrocs utilisaient les réseaux sociaux traditionnels, tels que les réunions de groupe hebdomadaires, pour établir progressivement des liens solides avec leurs victimes potentielles. Aujourd'hui, selon la North American Securities Administrators Association (NASAA) [64], avec la popularité croissante des réseaux en

ligne, les escrocs se sont tournés vers les réseaux sociaux pour établir des liens de confiance avec leurs victimes, puis tirer parti des données personnelles publiées dans les profils en ligne de ces dernières.

2.6.2 Menaces modernes

Les menaces modernes sont généralement propres aux environnements des réseaux sociaux en ligne. En général, ces menaces ciblent spécifiquement les informations personnelles des utilisateurs ainsi que celles de leurs amis. Par exemple, un attaquant qui tente d'accéder au nom de l'établissement scolaire d'un utilisateur de Facebook - visible uniquement par les amis Facebook de l'utilisateur - peut créer un faux profil avec des détails pertinents et lancer une demande d'ami à l'utilisateur ciblé. Si l'utilisateur accepte la demande d'ami, ses coordonnées seront exposées à l'attaquant. L'attaquant peut également recueillir des données auprès des amis Facebook de l'utilisateur et utiliser une attaque par inférence pour déduire le nom de l'école à partir des données recueillies auprès des amis de l'utilisateur. Nous citons dans ce qui suit neuf variétés de menaces modernes.

Clickjacking. Le clickjacking ou détournement de clics est une technique malveillante qui incite les utilisateurs à cliquer sur quelque chose de différent de sur quoi ils avaient l'intention de cliquer. En utilisant le clickjacking, l'attaquant peut manipuler l'utilisateur pour qu'il publie des messages de spam sur son mur Facebook, qu'il aime des liens à son insu (likejacking) et même qu'il ouvre un microphone et une webcam pour enregistrer l'utilisateur [65].

Attaques de désanonymisation. Les attaques de désanonymisation utilisent des techniques telles que le suivi des cookies, la topologie du réseau et l'appartenance à des groupes d'utilisateurs pour découvrir la véritable identité de l'utilisateur. Un exemple de désanonymisation a été démontré par Krishnamurthy et Wills [66], qui ont prouvé qu'il est possible pour des tiers de découvrir l'identité d'un utilisateur d'un réseau social en ligne en reliant les informations divulguées par les sites liés à ce réseau. Krishnamurthy et Wills ont également montré que la plupart des utilisateurs des réseaux sociaux étudiés étaient vulnérables à la fuite de leurs informations d'identité sur ces réseaux via des mécanismes de suivi, tels que les cookies de suivi.

Reconnaissance faciale. Les photos de profils d'utilisateurs des réseaux sociaux comme Facebook par exemple sont publiquement disponibles pour être visualisées et téléchargées. Ces photos peuvent être utilisées pour créer une base de données biométriques, qui peut ensuite être utilisée pour identifier par des algorithmes de reconnaissance faciale les utilisateurs du réseau en ligne sans leur consentement.

Faux-profils. Les faux-profils sont des profils automatiques ou semi-automatiques qui imitent les comportements humains dans les réseaux sociaux en ligne. Souvent, les faux profils sont utilisés pour récolter les données personnelles des utilisateurs sur les réseaux sociaux. En lançant des demandes d'amis à d'autres utilisateurs du réseau, qui acceptent souvent ces demandes, des robots sociaux peuvent recueillir les données privées d'un utilisateur qui ne devraient être exposées qu'à ses amis. De plus, les faux profils peuvent être utilisés pour lancer des attaques sybiles [67], publier des messages de spam [68], ou même manipuler les statistiques du réseau social en ligne [69][70].

Attaques par clonage d'identité. Grâce à cette technique, les attaquants reproduisent la présence en ligne d'un utilisateur, soit dans le même réseau, soit dans différents réseaux, afin de tromper les amis de l'utilisateur cloné et de les amener à établir une relation de confiance avec son profil. L'attaquant peut utiliser cette confiance pour collecter des informations personnelles sur les amis de l'utilisateur ou pour réaliser divers types de fraude en ligne.

Attaques par inférence. Les attaques par inférence dans les réseaux sociaux en ligne sont utilisées pour prédire les informations personnelles et sensibles d'un utilisateur que celui-ci n'a pas choisi de divulguer, telles que son appartenance religieuse ou son orientation sexuelle. Ces types d'attaques peuvent être mis en œuvre en utilisant des techniques d'exploration de données combinées avec des données du réseau accessibles au public, telles que la topologie du réseau et les données des amis des utilisateurs.

Fuite d'informations. Les réseaux sociaux en ligne permettent aux utilisateurs de partager et d'échanger ouvertement des informations avec leurs amis et les autres utilisateurs du réseau. Dans certains cas, les utilisateurs du réseau partagent volontairement des informations sensibles sur eux-mêmes et sur d'autres personnes, telles que des informations relatives à la santé [71][72] et au statut de sobriété [71]. La fuite d'informations sensibles et personnelles peut avoir des conséquences négatives pour les utilisateurs des réseaux sociaux. Par exemple, les compagnies d'assurance peuvent utiliser des données publiées sur des réseaux sociaux pour identifier les clients à risque [73]. En outre, les employeurs peuvent utiliser les réseaux sociaux pour sélectionner les candidats à l'emploi [74].

Fuite de localisation. Avec l'utilisation croissante d'appareils mobiles intelligents qui encouragent le partage d'informations de localisation [75], de nombreuses

personnes utilisent les réseaux sociaux en ligne pour partager volontairement des informations privées et parfois sensibles sur leur localisation actuelle ou future (ou celle de leurs amis). Dans certains cas, des utilisateurs du réseau partagent sans le savoir leur emplacement en téléchargeant des éléments médiatiques, tels que des photos et des vidéos, qui peuvent contenir des informations de géolocalisation sur leur emplacement actuel et passé [76].

Socware. Le socware consiste en des messages faux et éventuellement préjudiciables provenant d'amis dans les réseaux sociaux. Les socwares peuvent attirer leurs victimes en offrant de fausses récompenses aux utilisateurs qui installent des applications Facebook malveillantes liées aux socwares ou qui visitent des sites Web socwares douteux. Une fois le site Web du socware consulté ou l'application malveillante installée, le socware envoie des messages au nom de l'utilisateur à ses amis.

2.6.3 Menaces combinées

Les menaces classiques et modernes peuvent être combinées afin de créer une attaque plus sophistiquée. Par exemple, un attaquant peut utiliser une attaque par hameçonnage pour recueillir le mot de passe Facebook d'un utilisateur ciblé, puis publier un message contenant une attaque par clic sur le mur de l'utilisateur ciblé, incitant ainsi les amis Facebook de l'utilisateur à cliquer sur le message publié et à installer un virus caché sur leurs propres ordinateurs. Un autre exemple est l'utilisation de profils clonés pour collecter des informations personnelles sur les amis de l'utilisateur cloné, et en utilisant ces information, l'attaquant peut envoyer des messages de spam personnalisés contenant un virus. En utilisant les informations personnelles, le virus a plus de chances d'être activé.

2.6.4 Menaces visant les enfants

Certes, les enfants, jeunes ou adolescents, sont confrontés aux menaces classiques et modernes décrites ci-dessus, mais il existe également des menaces qui ciblent intentionnellement et spécifiquement cette catégorie d'âge dans les réseaux sociaux. Nous exposons dans ce qui suit trois variétés de cette catégorie de menaces, soient : la prédation en ligne, la cyberintimidation et le ciblage des comprtements à risque.

Prédation en ligne. La plus grande préoccupation concernant la sécurité des informations personnelles des enfants concerne les pédophiles sur Internet, également appelés prédateurs en ligne. Les comportements considérés comme de l'exploitation sexuelle d'enfants sur Internet comprennent l'utilisation d'enfants par des adultes pour la production de pornographie enfantine et sa distribution, la consommation de pornographie enfantine et l'utilisation d'Internet comme moyen d'initier une exploitation sexuelle en ligne. L'image du prédateur en ligne véhiculée par les médias est celle d'un homme adulte qui se fait passer pour l'ami d'un jeune garçon ou d'une jeune fille innocente par l'intermédiaire duquel il collecte des données personnelles ; il cache ses intentions sexuelles jusqu'à la rencontre réelle, qui implique probablement un viol ou un enlèvement. Dans la plupart des cas, les enfants sont conscients du fait qu'ils parlent à un adulte, et si la relation s'intensifie jusqu'à la participation à une rencontre réelle, ils sont conscients et, dans une certaine mesure, s'attendent à s'engager dans une activité sexuelle. Le plus souvent, la rencontre implique une activité sexuelle non forcée, mais elle a lieu avec une personne n'ayant pas l'âge du consentement et constitue donc un crime.

Cyberintimidation. La cyberintimidation est une intimidation qui a lieu sur des platesformes de communication technologiques, telles que les courriels, les

chats, les conversations téléphoniques et les réseaux sociaux en ligne, par un agresseur qui utilise la plateforme pour harceler sa victime en envoyant des messages blessants répétés, des expressions sexuelles ou des menaces, en publiant des photos ou des vidéos embarrassantes de la victime ou en adoptant d'autres comportements inappropriés. Aujourd'hui, la cyberintimidation est devenue un phénomène courant dans les réseaux sociaux dans lesquels l'attaquant peut utiliser l'infrastructure du réseau pour répandre des rumeurs cruelles sur la victime et partager des photos embarrassantes avec le réseau d'amis de la victime [77]. La cyberintimidation touche généralement les enfants, plutôt que les adultes.

Ciblage des comportements à risque. Les comportements à risque potentiels des enfants peuvent inclure la communication en ligne avec des inconnus, l'utilisation de salons de discussion pour des interactions avec des inconnus, des conversations sexuellement explicites avec des inconnus et la transmission d'informations et de photos privées à des inconnus. Il convient de noter que si chacun des comportements susmentionnés présente un risque à lui seul, la combinaison de quelques-uns de ces comportements peut, à juste titre, provoquer une énorme anxiété concernant la sécurité d'un enfant. En outre, il existe un lien bien établi entre les comportements en ligne et les comportements dans la réalié. En fait, plusieurs chercheurs affirment que les victimes d'abus sur Internet sont très souvent des enfants vulnérables, comme les jeunes ayant des antécédents d'abus physiques ou sexuels ou ceux qui souffrent de dépression ou de problèmes d'interaction sociale [78]. Tous les enfants qui vivent avec ce genre de problèmes courent un risque plus élevé d'être victimes d'abus sexuels sur Internet ou lors de rencontres initiées en ligne [78].

2.7 Conclusion

Dans de chapitre, nous avons définit d'où vient l'hétérogénéité des comportements des individus dans les réseaux sociaux. Tout d'abord, nous avons définit le comportement utilisateur en présentant les modèles informatiques qui le modélisent. Ensuite, nous nous sommes concentrés sur le comportement anormal des individus en décrivant leurs typologies. Par conséquent, nous avons désigné et argumenté notre choix des natures et des types d'anomalies que nous visons dans cette thèse. Enfin, nous avons présenté les catégories des menaces dans les réseaux sociaux en mentionnant leurs variétés. Dans le chapitre suivant, nous metterons la lumière sur la présentation des principales techniques d'analyse des réseaux sociaux. Etant motivés par la détection des comportements violent dans des structures complexes dynamiques, en particulier les comportements terroristes, nous nous appuyons sur l'utilisation de différentes techniques d'exploration de divers types de données et des algorithmes d'apprentissage automatique que nous présenterons dans le prochain chapitre.

Chapitre 3

Techniques d'analyse des réseaux sociaux

Sommaire

3.1 Introduction

Les plateformes sociales sont devenues des moyens incontournables d'échange des informations entre les individus dans le monde entier. Certaines personnes les utilisent pour faire valoir leur opinion sur les sujets de tendance alors que d'autres pour rester en contact avec leurs amis d'enfance et savoir ce qui se passe dans leur vie. L'ensemble des informations échangées et les comportements sur ces plateformes sont enregistrés instantanément et stockés sur des serveurs. Nous pouvons répertorier ces données en quatre familles différentes, à savoir : (1) les données publiques d'un utilisateur, (2) les données protégées par les paramètres de confidentialité (les discussions privées, la liste d'amis de l'utilisateur, etc.), (3) les données créées par d'autres membres (les identifications dans des publications, sur des photos, etc) et (4) les données créées directement par les réseaux sociaux en se basant sur les activités de l'utilisateur. Chaque famille de données contient divers types de données. Dans les sections suivantes, nous commençerons par donner un aperçu des types de données existants sur la structure des réseaux sociaux et de leurs approches d'analyse, puis nous présenterons les différents modèles de classification s'adaptant bien avec ces types de données.

3.2 Types de données

Dans [37], 270 fonctionnalités ont été identifiées pour décrire le comportement d'un individu, la plupart d'entre elles étant des fonctionnalités des médias sociaux. Nous nous sommes inspirés de cet ensemble de fonctionnalités pour trouver le moyen de les classifier en trois catégories selon leurs types, à savoir : les fonctionnalités textuelles, les fonctionnalités regroupant l'ensemble d'images et les fonctionnalités numériques.

3.2.1 Données textuelles

Les données textuelles sont principalement constituées de caractères appartenant à une langue spécifique et pouvant être lus par un être humain. Il s'agit principalement des publications, des commentaires, des biographies ou des légendes accompagnants des images. Il existe de nombreuses approches pour l'analyse de texte, principalement les techniques du Text Mining (TM) et du Natural Language Processing (NLP). Dans ce qui suit, nous présenterons ces deux techniques en expliquant l'importance de l'une par rapport à l'autre suite à notre objectif.

3.2.1.1 Analyse textuelle

Text Mining est le processus d'extraction d'informations de haute qualité à partir de données textuelles, l'information pouvant être des motifs ou des structures correspondantes dans le texte sans tenir compte de la sémantique de celui-ci. Il en résulte principalement des informations statistiques telles que la fréquence et la corrélation des mots [79].

Natural Language Processing est le processus qui permet à l'ordinateur de comprendre le langage parlé par les humains, ainsi que la sémantique et les sentiments qu'il transmet, en effectuant des analyses telles que l'analyse lexicale, syntaxique et sémantique [79].

Dans le domaine de détection des comportements atypiques, il est nécessaire d'extraire le contexte de l'analyse. Nous souhaitons savoir ce que l'utilisateur tente d'inciter avec ses publications et leur degré d'influence sur les autres utilisateurs du réseau. Par conséquent, nous devons passer par l'analyse sémantique et ne pas travailler avec les mots en tant qu'objets. Ainsi, le traitement automatique du langage naturel (NLP) semble être la technique la plus appropriée à notre besoin.

Le NLP est une branche multidisciplinaire qui combine des connaissances de

l'informatique, de la linguistique et de l'intelligence artificielle. Appelé également Traitement Automatique du Langage Naturel ou ingénierie linguistique, il vise principalement à analyser, traiter et reproduire le langage humain de manière automatique. Pour ce faire, il utilise des algorithmes spécifiques du Machine Learning (ML) et du DL. Grâce aux techniques du NLP, les machines peuvent interpréter et reproduire efficacement le langage parlé. Les différentes phases de NLP sont l'analyse morphologique, l'analyse syntaxique et l'analyse sémantique.

Analyse morphologique

Le traitement de la morphologie consiste à analyser la structure des mots. Nous étudions leur construction à partir d'unités significatives primitives appelées **morphèmes**. Cela nous aidera à diviser les différents mots/phrases d'un document en **jetons** qui seront utilisé lors d'une analyse ultérieure.

Morphèmes se sont les plus petites unités ayant une signification dans un mot. Il existe deux types de morphèmes, à savoir les "Lemmes (Stems)" et les "Affixes" ; les lemmes étant la base ou la racine d'un mot et les affixes pouvant être un préfixe, un infixe ou un suffixe qui ne semble jamais isolé, mais combiné à un lemme. En prenant l'exemple de la figure 3.1, nous pouvons voir comment nous divisons un mot en un lemme et des affixes.

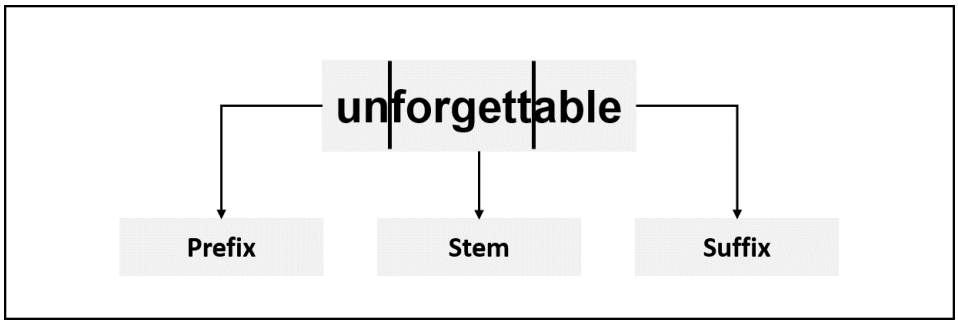

Figure 3.1: Exemple d'extraction de morphèmes.

Jetons se sont des mots, mots-clés, phrases ou symboles qui ont une unité sémantique utile pour le traitement. Le processus consacré à l'extraction de ces jetons est appelé "Tokenisation". Il est principalement composé d'un lemme + d'une partie d'une étiquette de parole + de caractéristiques grammaticales. A titre d'exemple :

— plays → play (lemma) + Noun (part of speech tag) + plural (grammatical feature)

— plays → play (lemma) + Verb (part of speech tag) + Singular (grammatical feature)

Les résultats de la phase morphologique sont utilisés par l'étape suivante, qui est l'analyse syntaxique.

Analyse syntaxique

L'analyse syntaxique consiste à mettre en évidence la structure d'un texte. Dans une phrase, la disposition des mots suit des règles précises de la grammaire de la langue. Par conséquent, cette analyse permettra à la machine de comprendre la relation entre les mots et les différentes références. En prenant un exemple de la phrase *"Two people were killed in an incident today"* et en suivant l'analyse syntaxique de la grammaire anglaise, nous aboutissons à l'exemple de la figure 3.2.

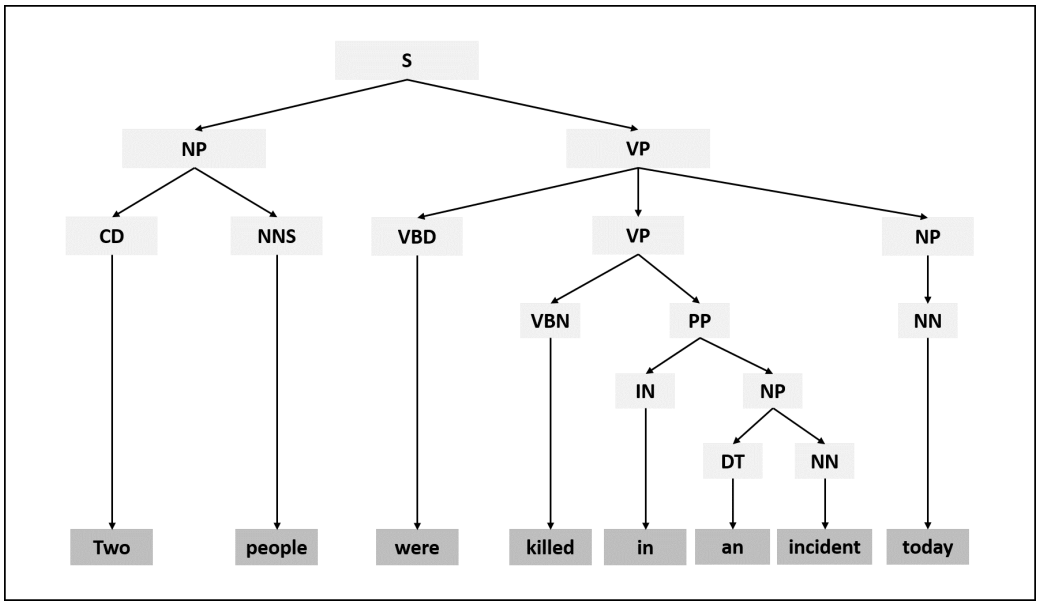

Figure 3.2: Exemple d'une analyse syntaxique.

Analyse sémantique

Après avoir structuré les mots et étudié leurs relations, la machine est prête à comprendre la sémantique des mots et des phrases, ainsi que le contexte du document. En se concentrant sur la relation entre les mots et certains autres éléments tels que les synonymes, les antonymes et les hyponymes (ordre de signification hiérarchique), le système sémantique peut construire des blocs composés de :

— Entités : individus ou instances.

— Concepts : Catégories d'individus ou de classes.

— Relations : relations entre entités et concepts.

— Prédicats : structures verbales ou rôles sémantiques.

Ces éléments peuvent être représentés à l'aide de différentes méthodes telles que First-order Predicate Logic (FOPL), les réseaux sémantiques et la dépendance conceptuelle. La figure 3.3 représente un exemple de réseau sémantique illustrant l'exemple traité à la phase précédente.

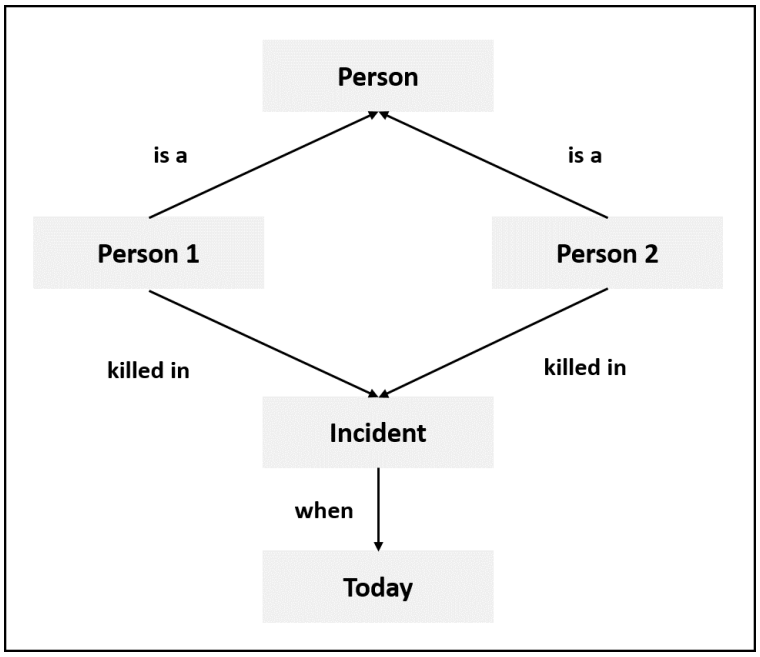

Figure 3.3: Exemple d'un réseau sémantique.

3.2.1.2 Représentation des données

L'application du NLP facilite à la machine la compréhension de la signification des données de contenu textuel. Mais afin de créer un classificateur qui catégorise automatiquement les données actuelles et futures, nos données doivent être converties en données numériques afin que nous puissions leur appliquer les méthodes mathématique, mais sans pour autant perdre la sémantique de ces données. L'incorporation de mots est l'une des représentations les plus courantes des données textuelles, qui consiste en la transformation d'un mot d'un document en un vecteur de caractéristiques numériques. La plupart des vecteurs proches signifie que ces mots partagent le même sens ou sont dans le même contexte, de sorte que les données ne perdent pas la sémantique. Lors de nos recherches, les techniques d'incorporation de mots les plus utilisées sont Word2Vec et Term Frequency-Inverse Document Frequency (TF-IDF). Nous présenterons, dans ce qui suit, ces deux

techniques en expliquant notre choix.

Word2Vec utilise deux approches différentes, à savoir : Continuous Bag Of Words (CBOW) et Skip Gram, toutes les deux sont basées sur des réseaux de neurones prenant un contexte en entrée et utilisant la technique de rétro-propagation pour apprendre [80]. Le travail mathématique d'arrière-plan de Word2Vec tentera de maximiser la probabilité du mot suivant w_t à partir du mot précédent h. L'équation 3.1 montre la façon de calculer la probabilité $P(w_t|h)$, où le $score(w_t, h)$ calcule la conformité entre w_t et le contexte h et $softmax$ est la fonction mathématique qui est une généralisation de la fonction logistique. CBOW apprend l'intégration d'un mot en le prédisant sur la base des mots qui l'entourent et qui sont considérés comme le contexte. Skip-Gram apprend l'intégration d'un mot en considérant le mot actuel comme contexte et en prédisant les mots qui l'entourent. Selon [80], Skip-Gram peut être meilleur avec peu de données et représente mieux les mots rares, alors que CBOW est plus rapide et représente mieux les mots fréquents.

$$P(w_t|h) = softmax(score(w_t, h)) \tag{3.1}$$

TF-IDF représente les mots avec des poids. Ces poids sont basés sur le produit de la fréquence déterminée multiplié par la fréquence inverse du document. En termes plus simples, les mots qui apparaissent beaucoup, mais partout, doivent recevoir très peu de pondération ou de signification. Ils ne fournissent pas une grande valeur. Toutefois, si un mot apparaît très peu ou apparaît fréquemment, mais seulement dans un ou deux endroits, il s'agit probablement d'un mot plus important qui doit être pondéré [81].

Term-Frequency (TF) est le pourcentage d'occurrence d'un terme t dans un document d. Comme illustré dans l'équation 3.2, nous calculons cela en prenant

le nombre de fois qu'un terme t apparaît dans le document d par le nombre total de mots dans le document d.

$$tf_{t,d} = \frac{n_{t,d}}{\sum_{term} n_{term,d}} \tag{3.2}$$

— $n_{t,d}$ représente le nombre d'occurrences de termes dans un document d.

— $\sum_{term} n_{term,d}$ représente la somme des occurrences de tous les termes apparaissant dans le document d qui correspond au nombre total de mots du document d.

Inverse-Document-Frequency (IDF) est le rang d'un terme t par sa pertinence dans un document d. L'équation 3.3 montre la formule mathématique permettant de calculer idf, ce qui fait en prenant le nombre total de documents N et en divisant par df_t le nombre de documents contenant le terme t.

$$idf(t) = log_e(\frac{N}{df_t}) \tag{3.3}$$

En utilisant TF-IDF, le poids $w_{t,d}$ du mot t dans un document d est obtenu comme indiqué dans l'équation 3.4 : en multipliant $tf_{t,d}$ par $idf(t)$.

$$w_{t,d} = tf_{t,d} * idf(t) \tag{3.4}$$

Dans notre recherche, nous avons examiné des travaux tels que celui de Shivangi [81], montrant que Word2Vec donne de meilleurs résultats en matière de gestion de mémoire et de temps d'exécution et intégre bien les mots de similarité en termes de contexte et de signification, alors que TF-IDF identifie mieux les mots qui déterminent la catégorie du document. Le tableau 3.1 résume les avantages et les inconvénients de chacune de ces deux techniques.

Table 3.1: Comparaison des méthodes d'incorporation de mots.

Méthode	Avantages	Inconvénients
Word2Vec	- Optimisation de l'utilisation de la mémoire - Rapidité du temps d'exécution	- Existance de beaucoup de données bruyantes - Non fonctionnement avec des données ambigues
TF-IDF	- Identification des catégories des données - Extraction des informations pertinentes	- Utilisation élevée de la mémoire - Non similarité des mots les plus proches dans leur sens mais dans leur catégorie du contexte du document

Suite à notre problématique, nous optons pour TF-IDF car lorsque nous traitons un problème de classification, nous sommes plus intéressés par la différenciation des catégories plutôt que par la représentation de la similarité des mots.

3.2.2 Données images

Les données de contenu images regroupent tout ce qui est une représentation visuelle, voire mentale, de quelque chose (objet, être vivant et/ou concept). Dans le contexte des réseaux sociaux, ce sont l'ensemble des images postées, envoyées, mises comme photo de profil, etc. Dans la littérature, différentes approches sont disponibles pour le traitement des images. Comme souligné dans [82], CNN est la méthode la plus performante dans la classification des images en termes de précision et de temps d'exécution. Ce qui différencie le CNN des autres méthodes est

la possibilité de considérer une structure spatiale avec une entrée multidimensionnelle au lieu de vecteurs aplatis et l'invariance de la traduction, ce qui signifie que, quel que soit le lieu où se trouve un objet dans une image, il est toujours considéré comme étant le même objet [83].

Dans cette section, nous expliquerons comment CNN fonctionne pour interpréter et classer les images, puis comment nous pouvons améliorer ses performances pour obtenir des meilleurs résultats.

3.2.2.1 Architecture CNN

Le réseau de neurones convolutionnel est un algorithme d'apprentissage en profondeur et une extension du réseau de neurones pouvant avoir une entrée multidimensionnelle contrairement aux réseaux de neurones classiques qui utilisent un vecteur comme entrée. Cet avantage lui confère de meilleures performances avec les données images car les images ont généralement trois canaux de couleur (RVB) représentables par une matrice à trois dimensions. Si nous prenons l'exemple d'une image de dimension 32×32 avec 3 canaux de couleur, nous aurions $32 \times 32 \times 3 = 3072$ poids pour un réseau de neurones régulier. Si nous choisissons une image de dimension 512×512, nous aurions $512 \times 512 \times 3 = 786432$ poids, ce qui donnera lieu à un calcul complexe ùénorme ainsi qu'à un sur-ajustement pour avoir beaucoup d'informations et de détails. Un autre avantage de CNN apparait dans l'application des filtres et la disposition d'un réseau de partage de poids, ce qui refléte leur fonctionnement avec les images car ce qui distingue une image d'une autre est sa structure spatiale [84].

Un simple CNN est composé d'un empilement de couches de traitement : une couche convolutionnelle, une couche de regroupement et une couche entièrement connectée. Dans un réseau CNN typique, il peut y avoir plusieurs tours de convolution/pooling jusqu'à arriver à la couche entièrement connectée.

Couche convolutionnelle

Chaque couche de convolution du réseau possède un ensemble de cartes de caractéristiques pouvant reconnaître de manière hiérarchique des motifs/formes de plus en plus complexes. Au lieu de multiplication régulières de matrice, la couche convolutionnelle utilise des calculs de convolution. Nous détaillons, dans ce qui suit, le fonctionnement de cette couche.

Pour détecter certains modèles dans une image, les filtres sont un moyen souvent utilisé. Ils offrent également une répartition du poids. L'exemple illustré dans la figure 3.4 montre un filtre qui détecte les bords incurvés d'une image.

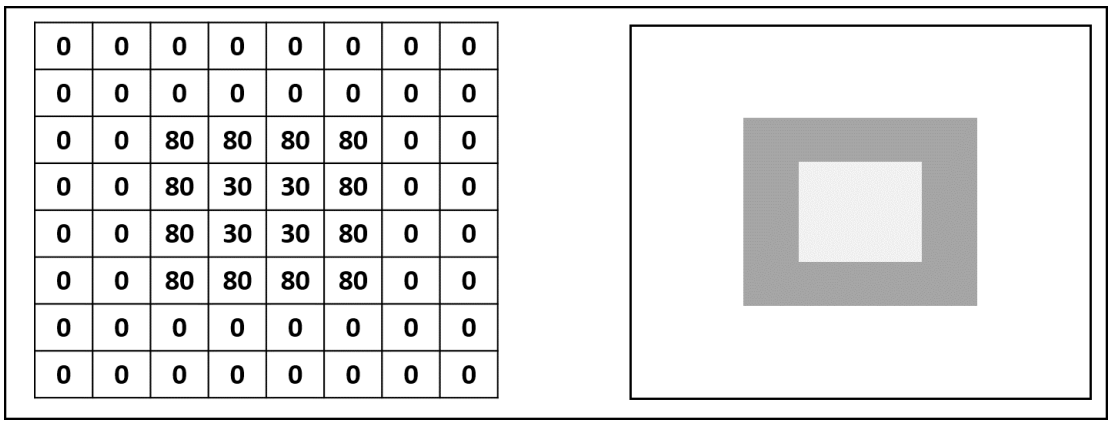

Figure 3.4: Filtre de bord incurvé.

L'identification d'un filtre nécessite le calcul suivant :

$$\begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 3 & 1 & 2 \\ 1 & 0 & 1 & 4 & 2 \\ 0 & 2 & 2 & 1 & 0 \\ 3 & 4 & 1 & 0 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \Downarrow & & \\ ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{pmatrix}$$

Pour obtenir la valeur du premier '?', nous devons utiliser le filtre sur la première

matrice de pixels 3×3 :? $= (0*1)+(0*1)+(1*0)+(1*0)+(1*0)+(3*1)+(1*1)+(0*0)+(1*0) = 4$.
Nous continuons, ensuite, le calcul de la valeur à côté de ' ?' qui est la valeur de la seconde matrice de pixels 3×3 dans laquelle '3' est le centre qui signifie que nous nous sommes déplacés de 1 pixel à droite.

$$\begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 3 & 1 & 2 \\ 1 & 0 & 1 & 4 & 2 \\ 0 & 2 & 2 & 1 & 0 \\ 3 & 4 & 1 & 0 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} & \Downarrow & \\ 4 & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{pmatrix}$$

$? = (0*1)+(1*1)+(1*0)+(1*0)+(3*0)+(1*1)+(0*1)+(1*0)+(4*0) = 2.$

Striding est un paramètre du nombre de pixels que nous allons déplacer pour calculer la prochaine valeur. Il est principalement utilisé pour réduire le calcul car la plupart des valeurs adjacentes ont plus de chance d'être similaires. Dans notre dernier exemple, le pas était égal à 1, ce qui signifie que nous n'avons déplacé que la case rouge de 1 pixel pour obtenir la valeur suivante. Habituellement, nous utilisons une valeur de 2 ou 3, car dans la plupart des cas, une distance de 2 à 3 pixels provoquerait une variation ou un changement de motif.

Padding est utilisé pour prévenir la perte d'informations. Dans notre exemple, lorsque nous appliquions le filtre, nous n'envisagions pas d'avoir les valeurs de la première/dernière ligne et de la première/dernière colonne au centre de la matrice 3×3. Pour corriger le problème, nous ajouterons de nouvelles lignes/colonnes remplies avec des 0.

$$\left(\begin{array}{|ccccc|}\hline 0&0&1&1&0\\ 1&1&3&1&2\\ 1&0&1&4&2\\ 0&2&2&1&0\\ 3&4&1&0&0\\ \hline\end{array}\right) \Longrightarrow \left(\begin{array}{ccccccc} 0&0&0&0&0&0&0\\ 0&\multicolumn{5}{|c|}{0\ 0\ 1\ 1\ 0}&0\\ 0&\multicolumn{5}{|c|}{1\ 1\ 3\ 1\ 2}&0\\ 0&\multicolumn{5}{|c|}{1\ 0\ 1\ 4\ 2}&0\\ 0&\multicolumn{5}{|c|}{0\ 2\ 2\ 1\ 0}&0\\ 0&\multicolumn{5}{|c|}{3\ 4\ 1\ 0\ 0}&0\\ 0&0&0&0&0&0&0\end{array}\right)$$

Couche Pooling

La couche de pooling consiste à réduire la taille des images, tout en préservant leurs caractéristiques importantes. Pour cela, nous découpons l'image en cellules régulières, puis nous gardons au sein de chaque cellule la valeur maximale. En pratique, nous utilisons souvent des cellules carrées de petites tailles pour ne pas perdre trop d'informations. Les choix les plus communs sont des cellules adjacentes de taille 2×2 pixels qui ne se chevauchent pas, ou des cellules de taille 3×3 pixels, distantes les unes des autres d'un pas de 2 pixels (qui se chevauchent). Nous obtenons en sortie le même nombre de caractéristiques maps qu'en entrée, mais celles-ci sont bien plus petites. Nous considérons un pooling 2×2 avec un pas de 2 pixels :

$$\left(\begin{array}{|cc|cc|}\hline 1&0&0&1\\ 3&2&0&2\\ \hline 0&0&4&2\\ 4&1&0&1\\ \hline\end{array}\right) \Longrightarrow \left(\begin{array}{|c|c|}\hline 3&2\\ \hline 4&4\\ \hline\end{array}\right)$$

Couche entièrement connectée

Une couche entièrement connectée est une couche sur laquelle toutes les entrées sont connectées à toutes les sorties. Dans un CNN, elle est utilisée pour déterminer finalement la classe qui sera assignée à l'entrée principale. Avant de passer à la couche entièrement connectée, il est nécessaire d'utiliser une technique appelée aplatissement afin de générer un vecteur nécessaire à cette couche. Le principe de

cette technique est de transformer chaque matrice 2D de pixels en 1 colonne de pixels.

3.2.2.2 Techniques d'optimisation

Après la construction du CNN, il est primordial d'appliquer de nombreux réglages comme l'ajout/la suppression de blocs de convolution afin d'ajuster l'architecture conceptualisée au type de la problèmatique traitée. D'une manière générale, pour trouver la meilleure architecture, il suffit de continuer à réentrainer le CNN. Cette tâche paraît fastidieuse en terme de temps d'exécution et de mémoire. Pour remédier à ce problème, il existe une technique appelée Transfert Learning Transfer Learning (TL) qui permet d'obtenir de meilleurs résultats plus rapidement. Une autre limite très connue rencontrée habituellement dans la classification d'images réside dans le fait qu'il n'y a pas suffisamment de données variées ou que le nombre d'échantillons est petit. Une solution pertinente à cela se manifeste dans la technique Data Augmentation (DA).

Transfer Learning

L'apprentissage par transfert est une technique qui permet à un modèle de tirer parti des connaissances acquises en résolvant un autre problème similaire. À titre d'exemple, un modèle qui a appris à reconnaître les voitures, il est possible d'utiliser ses connaissances pour reconnaître les camions [85]. Il est prouvé dans [86] que l'apprentissage par transfert pourrait montrer une énorme amélioration des résultats d'apprentissage en termes de précision, de temps d'exécution et d'utilisation de mémoire.

Data Augmentation

Avec peu de données et peu de variations, cela conduit généralement à un goulot d'étranglement dans les modèles de réseaux de neurones qui nécessite des milliers

d'échantillons d'apprentissage avec de nombreuses variantes pour pouvoir généraliser l'apprentissage. Dans de tels cas, l'augmentation des données joue un rôle important, car c'est une technique permettant de générer plus de données. Ceci est fait en utilisant des fonctionnalités telles que la rotation, l'ajout de points bruyants, la redimension, la translation, etc. L'application d'une telle technique peut forcément aider à améliorer le score du modèle, comme indiqué dans [87].

3.2.3 Données numériques

Les données à contenu numérique sont les données basées sur des nombres pouvant être interprétés statistiquement. Ce type de données ne nécessite pas de pré-traitement, il peut donc être directement intégré dans un modèle. Les modèles pour ce type de données sont pour la plupart les modèles généraux d'apprentissage automatique statistique que nous présenterons dans la section suivante.

3.3 Techniques de détection d'anomalies

3.3.1 Social Network Analysis

Social Network Analysis (SNA) combine de nombreuses méthodes et techniques qui facilitent l'analyse des réseaux sociaux. SNA mesure et cartographie le flux des relations et les changements de relations entre les entités physiques (humains, groupes, organisations, etc.). Les techniques SNA peuvent être très utiles pour la détection d'utilisateurs malveillants à l'intérieur d'un réseau. Certaines de ces techniques sont présentées comme suit :

— Degré de centralité est défini comme le nombre de liens directs ou de connexions à un nœud [88]. Un nœud avec une valeur de centralité de degré plus élevé est souvent considéré comme un hub et une entité active dans le

réseau.

— Centralité d'intermédiation [89] est une mesure permettant d'identifier un nœud, qui agit comme un pont pour établir des connexions avec d'autres groupes ou communautés du réseau. Il peut être défini comme le nombre de chemins les plus courts entre toute paire passant par un nœud.

— Centralité de proximité mesure de la vitesse à laquelle il est possible d'atteindre un nœud à tous les autres nœuds du réseau [88]. Elle peut être définie comme la longueur moyenne de tous les chemins les plus courts entre un nœud et tous les autres nœuds du réseau.

— La centralité des vecteurs propres est une mesure d'importance relative en termes d'influence d'un nœud sur ses nœuds voisins dans le réseau. Il est généralement utilisé pour trouver le nœud le plus central du réseau à l'échelle mondiale.

3.3.2 Machine Learning : classification et régression

ML est un sous-ensemble du domaine de l'intelligence artificielle, qui permet à la machine d'acquérir automatiquement des connaissances par expérience sans être explicitement programmée. En suivant certaines statistiques et concepts mathématiques, la machine recherche des modèles dans les données fournies, les apprend et prend de meilleures décisions à l'avenir [90]. Plusieurs types de méthodes d'apprentissage existent en Machine Learning.

3.3.2.1 Types d'apprentissage

Il existe quatre familles de méthodes d'apprentissage automatique, à savoir :

— **Apprentissage supervisé** construit un modèle mathématique d'un ensemble de données qui contient à la fois les entrées et les sorties souhaitées.

Les données sont appelées données d'apprentissage et consistent en un ensemble d'exemples d'apprentissage. Chaque échantillon d'exemple d'apprentissage a une ou plusieurs entrées et une seule sortie souhaitée. La machine doit apprendre une fonction qui mappe les entrées aux sorties. Et en fonction des données d'entraînement et de test, la mesure des performances de la classification ou de la régression se fait à travers le calcul de la valeur de F-score qui est la moyenne harmonique entre la précision et le rappel de la classe contenant les anomalies.

— **Apprentissage non supervisé (clustering)** se concentre sur l'identification des points communs dans les données et réagit en fonction de la présence ou de l'absence de tels points communs dans chaque nouvelle donnée (couleurs, langue, démographie, etc.). Etant donné un échantillon de données sans la sortie, la machine doit apprendre une fonction qui catégorise ces échantillons en fonction des modèles appris.

— **Apprentissage semi-supervisé** est une technique d'apprentissage hybride entre l'apprentissage supervisé et non supervisé, où certaines données sont étiquetées et d'autres non étiquetées. Etant donné un petit nombre de données avec la sortie souhaitée (données étiquetées) et d'autres données sans sortie (données non étiquetées), la machine doit apprendre une fonction qui peut étiqueter les données non étiquetées en utilisant les connaissances apprises des données étiquetées.

— **Apprentissage par renforcement** concerne la manière dont les agents logiciels doivent entreprendre des actions dans un environnement afin de maximiser une certaine notion de récompense cumulative. Il est différent de l'apprentissage supervisé et non supervisé, où il est principalement utilisé dans les véhicules autonomes ou pour apprendre à jouer à un jeu contre un adversaire humain (Échecs, jeux vidéo). Etant donné un échantillon de

données, certaines actions et récompenses liées aux actions, la machine doit apprendre une fonction qui trouve les actions optimales pour obtenir des récompenses maximales.

3.3.2.2 Classificateurs existants

Dans l'apprentissage supervisé, la classification et la régression sont les problèmes courants de l'identification d'un ensemble de catégories appartenant à une nouvelle observation sur la base d'un ensemble de données d'apprentissage contenant des observations (étiquettes) dont l'appartenance à la catégorie est connue. Ceci est rendu possible grâce aux classificateurs. Certains de ces classificateurs sont expliqués dans ce qui suit :

Support Vector Machine (SVM) Un modèle de SVM est une représentation des données dans l'espace. Les exemples d'une même catégorie sont proches les uns des autres. Le groupe d'exemples d'une catégorie est séparé par un écart clair aussi large et aussi espacé que possible des exemples d'une autre catégorie. Les nouveaux exemples observés sont ensuite prédits comme faisant partie d'une catégorie basée sur le côté de l'écart dans lequel ils se situent [91].

Logistic Regression (LR) Logistic Regression est un modèle statistique qui analyse une donnée dans laquelle au moins une caractéristique pourrait déterminer le résultat. En utilisant une fonction logistique, LR essaie de modéliser une sortie binaire qui est mesurée avec une variable dichotomique. Étant donné que la sortie est binaire, elle ne peut être utilisée que pour des problèmes de classification binaire. Pour l'utiliser pour un problème multi-classes, N modèles de régression logistique doivent être entraînés, où N est le nombre de classes des types de données [92].

Neural Network (NN) Un réseau de neurones est un réseau dans lequel il existe plusieurs couches de perceptrons. Un perceptron est l'unité élémentaire d'un réseau de neurones artificiels qui a été introduit comme modèle de neurones biologiques en 1959 [93]. La sortie de chaque perceptron d'une couche est connectée à chaque perceptron d'une autre couche en tant qu'entrée, ce qui la rend connue sous le nom de couche entièrement connectée. Un réseau de neurones doit avoir une couche d'entrée, une couche de sortie et entre eux une couche cachée. Tout réseau de neurones avec plus d'une couche cachée est considéré comme un réseau de neurones profond [84].

Naive Bayes (NB) Les classificateurs NB font partie des classificateurs probabilistes basés sur l'application du théorème de Bayes. Ils font partie des modèles de réseau bayésien qui sont hautement évolutifs et peuvent obtenir des résultats optimaux avec un minimum de données d'entraînement.

3.4 Conclusion

Dans ce chapitre, nous avons étudié les techniques existantes nécessaires pour effectuer une classification sur divers types de données ; données de contenu textuel, données de contenu image et données de contenu numérique. Le choix du classificateur dépend de la quantité des données collectée et du type des données mis en question. Dans le chapitre suivant, nous détaillerons un état de l'art spécifique à la détection des anomalies dans les réseaux sociaux en proposant un ensemble de critères d'évaluation pour mettre la lumière sur les limites des travaux existants.

Chapitre 4

Méthodes de détection d'anomalies dans les réseaux sociaux : Etat de l'art

Sommaire

4.1 Introduction

En raison de l'émergence des sites des réseaux sociaux tels que Facebook, Instagram, etc., le nombre d'impacts négatifs des phénomènes d'agressivité et d'intimidation a augmenté de façon exponentielle. La détection d'anomalies est un problème d'une importance capitale qui a suscité l'intérêt des chercheurs depuis les années 2000. Ce problème est souvent réalisé, grâce au deep learning, à l'intelligence artificielle et aux statistiques. Plusieurs méthodes ont été consacrées à la résolution du problème de détection des comportements anormaux sur les réseaux sociaux, qui se déclinent en trois types différents : les méthodes structurelles qui reposent sur l'analyse de graphes de réseaux sociaux, les méthodes comportementales qui reposent sur l'extraction et l'analyse des activités des utilisateurs et les méthodes hybrides qui combinent les deux types de méthodes mentionnés ci-dessus. Ce chapitre passe en revue diverses méthodes d'exploration de données pour la détection d'anomalies afin de fournir une meilleure évaluation pouvant faciliter la compréhension de ce domaine.

4.2 Méthodes de détection d'anomalies dans les réseaux sociaux

Les chercheurs étudient la détection d'anomalies selon trois méthodes principales : les méthodes basées sur l'analyse des activités (méthodes comportementales), les méthodes basées sur les graphes (méthodes structurelles) et les méthodes hybrides. La première catégorie se concentre sur les modèles qui traitent du contenu des activités des utilisateurs sur les réseaux sociaux. La deuxième catégorie met l'accent sur l'exploration des propriétés structurelles des graphes de

réseaux sociaux modélisant les relations entre les différents utilisateurs. La troisième catégorie analyse les activités des utilisateurs dans la structure des graphes des réseaux sociaux.

4.2.1 Méthodes comportementales

Les méthodes de détection basées sur l'activité considèrent que les utilisateurs sont quelque peu indépendants les uns des autres. De plus, l'évolution du comportement d'un utilisateur étant dépendante de sa fréquentation, cela pourrait constituer une limitation sérieuse de ce type d'approche. Par conséquent, un individu est défini par ses propres activités et cela déterminerait si son comportement est anormal [94][95][96][97][98][99][100][37][36]. Ces activités peuvent être le nombre de messages reçus et envoyés, le contenu des messages, la durée de navigation ou le temps passé sur un événement, le nombre de partages et de likes, le détail d'un élément partagé, etc.

Dans [37], les auteurs ont présenté une enquête sur les méthodes de profilage des utilisateurs disponibles pour la détection d'anomalies, puis ils ont proposé leur propre modèle de détection d'anomalies. Ils ont montré les avantages et les inconvénients de chaque modèle du point de vue de la cybersécurité. Certains modèles utilisent le journal du système d'exploitation et l'historique du navigateur Web comme source de données, tandis que d'autres sont davantage axés sur les réseaux sociaux tels que Twitter et Facebook. Leurs analyses ont révélé que les modèles basés sur l'historique et les journaux sont plus limités et incohérents car il est difficile de confirmer que c'est le même et l'unique utilisateur qui utilise le système d'exploitation ou le navigateur Web. Cependant, les modèles liés aux réseaux sociaux sont plus précis car il s'agit d'approches basées sur des comptes privés qui incluent également l'interactivité des utilisateurs entre eux, ce qui conduit à de meilleurs résultats. Notons que les données utilisées lors de l'analyse doivent être explicites

car les réseaux de délinquance et de fraude sont généralement offusqués ou déguisés. Sur la base d'autres méthodes de source de données, du trafic réseau, de l'historique du navigateur, des réseaux sociaux, des informations démographiques provenant des ressources humaines, des journaux système contenant des données de surveillance, etc., les auteurs ont défini la représentation du profil d'un utilisateur avec un vecteur de sept catégories de caractéristiques principales, à savoir : les caractéristiques des intérêts des utilisateurs , les fonctionnalités de connaissances et de compétences, les fonctionnalités d'informations démographiques, les fonctionnalités d'intention, les fonctionnalités de comportement en ligne et hors ligne, les fonctionnalités d'activité sur les réseaux sociaux et les fonctionnalités de trafic réseau. Chaque catégorie de fonctionnalités contient des fonctionnalités et des sous-groupes de fonctionnalités, ce qui a finalement conduit à plus de 270 fonctionnalités principalement liées à la sécurité. Leur modèle proposé appelé "Unified User Profiling" collecte principalement les données des différentes sources, puis les nettoie et les analyse pour obtenir des données structurées permettant d'avoir un vecteur de profil d'utilisateur que l'administrateur est capable de surveiller dans différentes catégories et de détecter les anomalies en fonction de l'activité de l'utilisateur. Bien que leur modèle soit majoritairement complet en termes de fonctionnalités et de sources de données différentes par rapport à d'autres modèles qui n'utilisent que deux sources de données au maximum, il reste néanmoins limité car il ne détecte pas automatiquement les anomalies ; une intervention humaine est nécessaire.

Dans [36], les auteurs ont proposé une méthode de reconnaissance de formes, qui, étant donné un vecteur du profil d'un utilisateur, prend les activités quotidiennes de l'utilisateur et crée un modèle de série temporelle pour cet utilisateur sur chaque activité qu'il fait, puis chaque fois que l'utilisateur est impliqué dans une activité, le nouveau comportement est comparé à son modèle comportemental

de cette activité. Si un écart par rapport au comportement normal s'est produit, il est signalé comme suspect. Cependant, étant donné qu'un écart mineur ne signifie pas toujours un soupçon, il existe un modèle comportemental de tous les utilisateurs du système auquel l'activité sera également comparée, afin que les fausses alarmes soient réduites au minimum. Leur modèle est une forêt aléatoire formée sur l'ensemble de données CERT [101] ainsi qu'un ensemble de données privé acquis auprès de NextLabs qui a atteint une précision de plus de 97%. Cette précision est due au bon choix de l'algorithme de forêt aléatoire. Bien que cette méthode ait donné d'excellents résultats en termes de détection des menaces internes, on peut identifier deux limites majeures. Premièrement, le modèle construit ne supporte pas les réseaux multiples car il n'y a pas de relations entre les différentes entités analysées, ce qui implique l'impossibilité de modéliser le comportement des utilisateurs par un graphe monodimensionnel ou multidimensionnel. Deuxièmement, le modèle est incapable d'apprendre automatiquement les comportements futurs au fil du temps.

Le problème de classification OpenWorld est très souvent présent dans des environnements ouverts dynamiques (par exemple, les réseaux sociaux) où certains nouveaux éléments peuvent n'appartenir à aucune des classes d'apprentissage existantes. Shu et al. [98] ont construit un modèle qui peut classer les nouveaux documents texte dans des classes déjà connues ou les rejeter pour montrer qu'ils ne sont conformes à aucun des modèles représentant les classes existantes. L'idée de ce classificateur est d'avoir des classes "m + 1" où les classes "m" sont pour l'apprentissage et une classe pour le rejet. La méthode proposée est appelée Deep Open Classification (DOC), et est principalement basée sur l'apprentissage en profondeur et utilise un réseau de neurones de type Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM). DOC est testé sur un ensemble de données [102] qui contient des critiques Amazon de 50 classes de produits où chaque

classe avait 1000 critiques. Bien que les résultats obtenus dépassent largement les approches existantes de la classification des textes, ce classificateur ne considère pas les données multimodales. À notre avis, les auteurs n'ont pas considéré l'idée d'utiliser divers types de données d'entrée bien que leur classificateur soit construit sur la base de l'architecture CNN-LSTM, qui peut prendre en charge des données multimodales.

Analyser la crédibilité des rumeurs issues des réseaux sociaux est un domaine de recherche d'actualité. Ces fausses informations se propagent plus rapidement et plus longtemps que les vraies informations. Par conséquent, leur structure de diffusion est différente de celle de l'actualité réelle [103]. Les utilisateurs de médias sociaux se comportent de manière très remarquable lorsqu'ils publient des rumeurs et des messages normaux. Dans ce contexte, Zhang et al. [99] ont proposé un modèle basé sur un AE à deux couches (autoencodeur). Les auteurs ont commencé par modéliser le comportement normal des utilisateurs à partir de leurs récentes publications collectées par Sina Weibo [1]. Cette modélisation Weibo est basée sur 14 Fonctionnalités, à savoir : Nombre d'utilisateurs qui ont favorisé ce Weibo, Nombre d'utilisateurs qui l'ont commenté, Nombre d'utilisateurs qui l'ont reposté, Score de sentiment du Weibo, Nombre de photos postées, de sujets, de @mentions, de smileys, de points d'interrogation, et de pronoms à la première personne dans ce Weibo, Longueur du Weibo, Si le Weibo est une rediffusion, Heure où le Weibo a été publié et Comment le Weibo a été publié. Les résultats des tests ont atteint une précision de 88% sur un ensemble de données contenant 167731 de Weibos. Les autoencodeurs ne posent pas de problème en termes d'utilisation de données multimodales. Par conséquent, les données non prises en compte (e.g., images, informations personnelles, etc.) sont absentes dans toutes les données de test. Dans ce cas, la limitation ne vient pas du modèle lui-même mais de l'indisponibilité des

1. Weibo Community Management Center for false rumor category : http ://service.account.Weibo.com/ ?type=5.

données.

Malgré la nécessité de méthodes de détection de contenus indésirables de vidéos sur les réseaux sociaux, peu de travaux s'intéressent à ce sujet. Ceci s'explique par le fait que les analyses d'images et de vidéos reposent sur des techniques moins variées que celles des autres médias (ex : segmentation d'images, détection de contours, etc.). Dans ce contexte, Yang et al. [100] ont développé un Weighted Convolutional Autoencoder (WCAE) réseau qui utilise des autoencodeurs pour détecter les variantes spatiales et 3 convLSTM pour capturer des variantes temporelles complexes. Ce réseau a été testé sur les deux jeux de données Chinese Univeristy of Hong Kong (CUHK) [2] [104] et University of California San Diego (UCSD) [3] [105] qui contiennent des déplacements dans les avenues. Toutes les vidéos de la partie apprentissage ont enregistré des événements normaux, tels que la marche régulière dans les allées, et les vidéos de la partie de test incluaient des mouvements irréguliers, tels que patiner, courir, lancer des objets, etc. Une comparaison avec les approches de détection d'anomalies basées sur les fonctionnalités ou l'apprentissage en profondeur montrent la supériorité du réseau construit avec une précision de 85%. Ce résultat explique le fait que la détection de comportements aberrants dans les flux vidéo est dû au domaine de la sécurité et donc concerne moins les réseaux sociaux. Cette méthode peut être améliorée en ayant recours à des tests sur des jeux de données plus enrichis avec des informations supplémentaires telles que les détails des vidéos, l'identité des personnes qui publient ces vidéos, etc.

Dans [28], les auteurs ont mis en œuvre un modèle qui détecte les extrémistes dans les médias sociaux sur la base de certaines informations liées aux noms d'utilisateur, aux profils et au contenu textuel. Ils ont construit leur ensemble de données

2. Abnormal Event Detection at 1000 FPS in MATLAB : http ://www.cse.cuhk.edu.hk/ leojia/projects/detectabnormal/index.html.

3. Privacy Preserving Crowd Monitoring : Counting People without People Models or Tracking : http ://www.svcl.ucsd.edu/projects/peoplecnt/.

à partir de Twitter en recherchant des hashtags liés à l'extrémisme, ce qui se traduit par environ 1,5M de tweets. De ce fait, ils ont extrait 150 comptes liés à Islamic State of Iraq and Syria (ISIS) [4] où les tweets de ces comptes ont été signalés au compte Twitter Safety (@TwitterSafety) comme étant des utilisateurs anormaux. Ils ont également extrait 150 comptes d'utilisateurs normaux afin d'avoir un ensemble de données équilibré. Ils ont augmenté le volume des données collectées par environ 3k de données non étiquetées. Ensuite, ils ont classé les fonctionnalités en trois groupes principaux : le nom d'utilisateur de Twitter, le profil et les fonctionnalités liées au contenu. Sur la base de cet ensemble de données, ils ont posé deux questions de recherche : les extrémistes sur Twitter sont-ils enclins à adopter des méthodes similaires ? Est-il possible de déduire les labels (extrémistes vs non-extrémistes) des comportements invisibles en fonction de leur proximité avec les instances étiquetées ? Après plusieurs expériences avec des approches supervisées et semi-supervisées, les deux questions ont eu des réponses positives. Le Support Vector Machine (SVM) a obtenu le meilleur score de précision (96%) tandis que le Char-Long Short-Term Memory (Char-LSTM) a obtenu le meilleur score de rappel de précision (76%) et réduit ainsi le nombre de faux négatifs. Un autre travail dans [28] a présenté différentes manières de collecter les données nécessaires à la détection des extrémistes. Les auteurs ont également montré que l'utilisation de différents types de données d'entrée provenant des réseaux sociaux est utile. La limitation de ce modèle est qu'il ne prend pas en charge les changements de comportement des utilisateurs au fil du temps et qu'il ne peut pas prédire les futurs comportements extrémistes.

Dans [29], les auteurs ont présenté un Convolutional Neural Network (CNN) afin de détecter les e-crimes suspects et l'implication terroriste en classant le

4. https ://www.kaggle.com/fifthtribe/how-isis-uses-twitter

contenu des images des réseaux sociaux. Ils ont utilisé trois types différents d'ensembles de données. Sur la base de la technique TL, ils ont pris l'architecture CNN du modèle ImageNet [82] et ils ont réduit la taille de son réseau en abaissant la taille du noyau de chaque couche pour créer leur nouveau réseau plus petit. Dans les résultats, leur architecture a surpassé l'ImageNet par défaut d'environ 1% du score de précision moyen et a pris la moitié du temps d'exécution d'ImageNet. Ce travail a montré que le concept de détection des terroristes sur la base du contenu de leurs images de réseaux sociaux est possible avec l'avantage d'utiliser TL plutôt que de construire un CNN à partir de zéro. Néanmoins, leur modèle ne supporte qu'un seul type de données, les images.

Dans [106], les auteurs ont développé une approche manuelle pour détecter les individus plus enclins à l'extrémisme. Leur démarche a commencé par explorer les réseaux sociaux afin de collecter les données de douze terroristes connus sur Twitter (3542 comptes ont été collectés et un maximum de 200 tweets par compte a été analysé) tout en mesurant trois dimensions pour chaque utilisateur : (1) leur influence : nombre de fois que leur contenu a été retweeté, (2) leur exposition : nombre de fois qu'il a retweeté le contenu des autres et (3) leur interactivité : en recherchant des mots-clés dans les tweets. Ce travail a bien montré que les scores élevés d'influence et d'exposition montraient une forte corrélation avec l'engagement de l'idéologie terroriste.

Dans [107], les auteurs ont créé un instantané démographique des partisans de ISIS sur Twitter en décrivant une méthodologie pour détecter les comptes pro-ISIS. À partir d'un ensemble de 454 comptes (identifiés par des recherches précédentes [106]), ils ont obtenu une liste finale de 20k comptes pro-ISIS à analyser. Ils ont estimé qu'au moins 46k comptes pro-ISIS étaient actifs (décembre 2014). Ce travail a créé des classificateurs à partir d'un sous-ensemble de 6k comptes annotés manuellement en tant que partisan ou non-partisan de l'état islamique.

Les auteurs ont conclu que les partisans pro-ISIS pouvaient être identifiés à partir de leurs descriptions de profil : avec des termes tels que *succession*, *linger*, *Islamic State*, *Caliphate State* ou *In Iraq all being prominent*. En testant ce classificateur avec 1574 comptes annotés manuellement, ils ont obtenu un score de 94% de précision.

Agarwal et Sureka [89] ont étudié des techniques pour identifier automatiquement les tweets encourageant la haine et l'extrémisme. À partir de deux analyses de données Twitter, ils ont utilisé une approche d'apprentissage semi-supervisée basée sur une liste de hashtags *(#Terrorism, #Islamophobia, #Extremist)* pour filtrer les tweets liés à la haine et à l'extrémisme. L'ensemble de données d'entraînement contient 10486 tweets. De plus, ils ont utilisé un échantillonnage aléatoire pour générer l'ensemble de données de validation (1M de tweets). Les tweets étaient en anglais et annotés manuellement. Ce travail a utilisé deux classificateurs différents (K-Nearest Neighbors (KNN) et SVM) sur la base des ensembles de données générés pour classer un tweet comme incitant à la haine ou inconnu. En validant ces classificateurs, les auteurs ont conclu que la présence de termes religieux, liés à la guerre, de mots offensants et d'émotions négatives sont des indicateurs forts d'un tweet suspect.

La détection du terrorisme est généralement considérée comme un problème binaire plutôt que comme un processus à différents degrés ou niveaux. La détection n'intervient que sur quelques utilisateurs sans tenir compte de la structure de leur réseau. Les approches existantes catégorisent les utilisateurs sur la base d'une petite quantité d'informations textuelles (commentaires, messages, etc.) mais peu de travaux prennent en compte les données d'images fournies par les utilisateurs.

4.2.2 Méthodes structurelles

Les méthodes de détection basées sur des graphes prennent en compte l'interactivité de l'utilisateur en analysant un instantané d'un réseau. Chaque utilisateur peut avoir une relation avec d'autres utilisateurs à travers des mentions, des partages et des likes. Cela peut être fait de manière statique ou dynamique. Dans les méthodes de détection basées sur des graphes statiques [27][108][22][24][9][25][109], l'analyse est effectuée sur un seul instantané du réseau, tandis que pour les méthodes de détection basées sur des graphes dynamiques [23][26][110], l'analyse est effectuée en un temps basé sur l'analyse d'une série d'instantanés.

Dans Akoglu et al. [24], l'algorithme OddBall présente une méthode rapide et non supervisée pour détecter les nœuds anormaux dans les graphes pondérés en mentionnant les règles appropriées à éliminer avant de classer un nœud comme anomalie. Cet algorithme détecte l'écart du comportement anormal par rapport à un comportement normal connu. La question qui se pose à cet effet est : qu'est-ce qu'un comportement normal connu ? Un comportement normal en 2019 peut ne pas être un comportement normal en 1970. L'évolution du même comportement dans le temps ne montre pas l'efficacité d'OddBall d'autant plus que les graphes testés sur cet algorithme ne sont pas des graphes d'évolution temporelle.

Hassanzadeh et al. [9] ont préconisé un nouveau framework basé sur le calcul d'un certain nombre de mesures (ego, egonet, super egonet, centrality, community, etc.) d'un graphe. Ce framework vise à détecter l'apparence générale d'un modèle suivi par la plupart des nœuds, puis il calcule un score aberrant de chaque nœud en fonction de la distance de la ligne d'ajustement pour distinguer les utilisateurs qui peuvent être anormaux et enfin il calcule un seuil pour minimiser le nombre de faux négatifs et le taux de faux positifs. D'une part, ce travail traite le fait que les réseaux sociaux ont une structure communautaire. Cela prouve que la majorité des utilisateurs appartiennent à un petit nombre de communautés. D'autre part, les

utilisateurs ayant un comportement anormal ont recours à l'établissement de relations aléatoires avec des utilisateurs appartenant à des communautés différentes. Cette méthode a été appliquée à des ensembles de données statiques provenant de réseaux sociaux en ligne.

Rezaei et al. [25] ont proposé une méthodologie basée sur le calcul de métriques de graphes. Cette méthodologie utilise la formule de mesure des anomalies Odd-Ball [24], qui est un moyen efficace de détecter un comportement anormal. Ce travail a réussi à marquer un ensemble important de nœuds avec une forte probabilité que ces nœuds puissent suivre un modèle anormal. A cet effet, les résultats obtenus ont prouvé que les comportements anormaux ont un petit nombre d'amis communs avec leurs amis. Les contraintes de temps manquent pour dynamiser cette méthode.

Fire et al. [22] ont proposé un algorithme pour la détection des spammeurs et des faux profils dans les réseaux sociaux. Cet algorithme prend en considération les communautés construites suite aux relations entre les utilisateurs. Il suppose que les utilisateurs anormaux se connectent au hasard à d'autres utilisateurs appartenant à des communautés différentes. Suffit-il que cette solution repose uniquement sur l'analyse de la topologie de la structure des réseaux sociaux ? L'évaluation de l'algorithme repose sur son exécution sur différentes structures de graphes statiques de réseaux sociaux.

Zheleva et al. [23] ont présenté un cadre pour la prédiction du type de relations entre les utilisateurs de réseaux sociaux. Ce travail s'appuie sur la combinaison de réseaux sociaux et de liens d'affiliation pour instancier des sous-graphes d'amitié et de famille dont le but est d'identifier des anomalies dynamiques anticipant des événements futurs. Les résultats de trois sites de médias sociaux ont validé l'efficacité du cadre proposé. Cette méthode ne peut être appliquée qu'aux jeux de données où les liens vers les groupes sont prédéfinis.

Chen et al. [26] ont développé un algorithme de détection d'anomalies basé sur la construction de communautés dans un réseau dynamique. Cet algorithme évolutif a montré l'efficacité des communautés dans les réseaux dynamiques par rapport à un algorithme non représentatif sur les réseaux statiques. À cette fin, les communautés sont la métrique la plus importante extraite d'un graphique. Cette approche a permis la détection de six types d'anomalies communautaires existantes dans les réseaux évolutifs, mais malheureusement cette solution ne peut pas être appliquée sur des réseaux complexes multidimensionnels.

Li et al. [110] ont examiné l'hypothèse que s'il y a un changement anormal dans la structure interne d'un réseau organisationnel, cela conduit à des événements externes et lorsque ces événements externes se produisent, la structure interne du réseau devient anormale par rapport à ses modèles normaux. En conséquence, les auteurs ont proposé un cadre axé sur la détection de comportements anormaux dans les réseaux organisationnels. Ce travail est basé sur l'utilisation d'un Back Propagation Neural Network (BPNN) utilisant deux méthodes heuristiques : les Genetic Algorithm (GA) et le Particle Swarm Optimization (PSO). Dans la phase de test, le Framework a été testé sur des réseaux longitudinaux construits à partir de la base de données Enron [5] [111] et 60 réseaux longitudinaux construits à partir de la base de données John Jay & ARTIS Transnational Terrorism Database (JJATT) [6] [112] qui ont enregistré les membres de l'Al- réseau Qaida et les liens opérationnels entre eux. Les résultats expérimentaux ont montré que les BPNN optimisés sont meilleurs que les BPNN simples. Ces dernières sont également meilleures que les méthodes de détection d'anomalies non supervisées telles que le Statistical Process Control (SPC), le Local Outlier Factor (LOF), le SVM

5. Enron Email Dataset http ://www.cs.cmu.edu/ enron/
6. John Jay & ARTIS Transnational Terrorism Database : http ://doi-tapps.jjay.cuny.edu/jjatt/index.php

à une classe One-Class SVM (OC-SVM), etc. et les méthodes de classification supervisées telles que la Logistic Regression (LR), Decision Trees (DT) et SVM. En conclusion, nous constatons que les auteurs auraient dû utiliser des données de différents types au cours de la phase de test et auraient dû essayer de synchroniser les réseaux pour extraire plus de connaissances, car les techniques de base de leur cadre supportent ces deux suggestions. Le problème revient donc à l'absence de données multimodales et à la manière de modéliser les réseaux.

Dans [109], Castellini et al. ont implémenté un Denoising Autoencoders (DAE) comme détecteur d'anomalies sur la plateforme Twitter formé avec une approche d'apprentissage semi-supervisé. Ce type d'approche ne nécessite pas le prélèvement d'échantillons anormaux. Les auteurs ont eu recours à l'identification des profils normaux par 17 caractéristiques telles que l'âge du compte, le nombre de tweets, le nombre d'amis, etc. Ce DAE a été testé sur un ensemble de données contenant des données extraites du CNR italien [7] et des données achetées sur différents marchés en ligne [8, 9]. Bien que ce détecteur traite le problème de détection d'anomalie comme un problème de classification binaire, il ne prend pas en compte la multimodalité des données due au mauvais choix de type de données lors de la phase de test.

En plus de ces études, les auteurs de [31] ont proposé une enquête sur l'analyse des réseaux sociaux pour lutter contre le terrorisme. Il ont fournit des méthodes de collecte de données ainsi que plusieurs types d'analyse. Deux principales sources de données sont identifiées : les réseaux sociaux en ligne (par exemple, Facebook, Twitter ou YouTube) et les réseaux sociaux hors ligne (par exemple, les bases de données publiques telles que Global Terrorism Database (GTD) [113] et Global Database of Events, Language, and Tone (GDELT) [114]). En conclusion, ils se

7. The Fake Project : http ://wafi.iit.cnr.it/theFakeProject/.
8. Friendly note : https ://www.fastfollowerz.com/closed/.
9. Twitterboost.com : http ://twitterboost.com/.

sont retrouvés avec l'idée que lors de l'exécution d'une analyse de réseau social, le principal défi repose sur les données elles-mêmes, car la confidentialité des utilisateurs est une question très sensible. Généralement, les informations manipulées sont incomplètes, suite par exemple au masquage de certaines relations, ce qui conduit souvent à des résultats d'analyse incorrects ou non précis.

Les terroristes adoptent un comportement dynamique au fil du temps. Par conséquent, il est très difficile de prédire leurs événements futurs car parfois aucun signe n'est détecté [115][116][117]. Dans [118], les auteurs ont recommandé une solution à ce problème. Ils ont suggéré un cadre pour détecter différents modèles de mouvements terroristes. Ces modèles définissent les motivations des attentats ainsi que les relations des acteurs des activités terroristes. Ce nouveau Cadre a permis d'analyser et de déduire de nouvelles connaissances sur les futures actions envisagées par les groupes extrémistes. Les expériences ont prouvé l'efficacité de ce cadre avec une précision de plus de 90%. Ces travaux visent principalement la détection et la prédiction d'événements terroristes. Elle se limite à la collecte de données textuelles et à l'analyse d'un réseau de relations unidimensionnel.

Mahmood et Alanezi [119] ont développé une nouvelle approche pour détecter les comportements anormaux potentiels sur Facebook. L'idée principale de ce travail est de combiner les caractéristiques structurelles et spectrales afin de suivre et révéler les anomalies au sein des interactions lors des manifestations irakiennes d'octobre 2019. Cette combinaison a permis de montrer des résultats expérimentaux efficaces en termes de détection d'événements anormaux. Deux limites principales de cette approche peuvent être discutées. Les auteurs ont dû extraire les mêmes données d'événements d'autres réseaux sociaux que Facebook pour enrichir leur base de connaissances. De plus, ils devaient tenir compte des propriétés comportementales pour s'assurer que leur processus de détection était précis et complet.

La détection d'anomalies dans les réseaux d'information multidimensionnels est un domaine de recherche actuel. Bien que les travaux présentés précédemment étudient bien la détection d'anomalies dans les réseaux unidimensionnels, seuls deux travaux récents ont abordé ce sujet dans les réseaux multidimensionnels [27][120]. Ce type particulier de réseau se caractérise par la synchronisation des réseaux étudiés. Cette synchronisation présente deux avantages principaux, à savoir : la richesse de la communication et l'augmentation des échantillons de données.

Dans [27], les auteurs ont développé une approche qui traitait l'hypothèse qu'un nœud atypique est un nœud faiblement connecté aux autres nœuds du réseau, dans toutes les dimensions d'un réseau multidimensionnel. Pour éviter de choisir un mauvais seuil pour la classification des scores obtenus pour chaque nœud, la distribution Beta a été utilisée. Chouchane et al. [27] ont généré cinq réseaux synthétiques avec différents paramètres tels que le nombre de dimensions pertinentes et non pertinentes et le nombre de nœuds normaux et anormaux pour évaluer l'efficacité de leur méthode. Les résultats obtenus ont montré que l'approche a identifié avec 95% de précision les anomalies même avec la présence d'un grand nombre de dimensions non pertinentes. Cette solution est limitée aux réseaux statiques et ne prend pas en compte la structure du réseau, ce qui oblige la construction de communautés.

Dans [120], les auteurs ont proposé un framework qui utilise un réseau multidimensionnel comme entrée pour l'identification des acteurs clés du réseau terroriste. Les dimensions représentent les types de relations ou d'interactions dans le réseau social. La méthodologie de leur cadre commence par la construction d'un réseau multidimensionnel via une recherche par mot-clé sur une plate-forme de médias sociaux, puis ce réseau est mappé sur un réseau à couche unique à l'aide de certaines fonctions de mappage. Pour détecter les acteurs clés, ils ont utilisé plusieurs

mesures de centralité telles que la centralité de degré et la centralité d'intermédiarité. Le résultat du cadre est une liste classée des acteurs clés au sein du réseau. L'efficacité du cadre a été évaluée avec un ensemble de données de vérité terrain d'une période de 16 mois de données Twitter. Ce travail a présenté l'utilisation de réseaux multidimensionnels pour la détection d'acteurs terroristes clés. Leur utilisation des dimensions multiples pourrait être plus efficace s'ils considéraient plusieurs réseaux sociaux au lieu de plusieurs types de relations et d'interactions.

4.2.3 Méthodes hybrides

Les méthodes hybrides interviennent explicitement et simultanément dans l'analyse des relations d'un utilisateur dans la structure de son réseau et de ses propres activités [121][35].

Dans [121], les auteurs ont suggéré un cadre de détection d'anomalies selon lequel, à chaque horodatage t, chaque utilisateur d'un réseau a un score d'activité et un score mutuel avec les autres utilisateurs. Les scores sont basés sur les activités de l'utilisateur et l'interactivité avec d'autres utilisateurs sur ces activités. Une matrice d'accord mutuel est ensuite produite pour représenter les scores où les activités de l'utilisateur sont des scores marqués dans la matrice diagonale. À l'aide d'une fonction de notation des anomalies qu'ils ont proposée, les scores de l'utilisateur y sont transmis et seuillés pour définir si l'utilisateur est anormal ou non. Ils ont utilisé le "CMU CERT Insider Threat Dataset" [10] [101] et le "NATOPS Gesture Dataset" [11] [122] comme sources de données, puis ils ont comparé les résultats de leur framework à d'autres modèles connus. Leur modèle était de loin le meilleur ; ils ont atteint environ 0,95 du score de zone sous la courbe, tandis que les autres modèles tels que SVM et le clustering étaient d'environ 0,89. Malgré

10. Insider Threat Test Dataset : https ://resources.sei.cmu.edu/library/asset-view.cfm ?assetid=508099.
11. NATOPS Aircraft Handling Signals Database : http ://groups.csail.mit.edu/mug/natops/.

le fait que ce framework a dépassé de loin les résultats attendus pour la détection des menaces internes et sa capacité à prendre en charge les changements de comportement au fil du temps, sa fonction de notation n'a pas pris en charge le scénario d'analyse d'un graphe multidimensionnel où chaque utilisateur peut avoir plus qu'un vecteur d'activités.

Dans [35], les auteurs ont proposé une approche de profilage de l'utilisateur basée sur ses caractéristiques comportementales et les caractéristiques des connexions aux réseaux sociaux. Le premier ensemble de fonctionnalités est le fondement de la représentation de l'utilisateur, qui est composé de statistiques sur le contenu des publications, de la sémantique du contenu des publications et des statistiques sur le comportement de l'utilisateur. Les fonctionnalités des connexions aux réseaux sociaux sont essentiellement un ensemble de fonctionnalités qui conduisent à la construction d'un réseau d'utilisateurs similaires qui ont des représentations de réseau similaires. Les résultats de l'expérience ont montré qu'en utilisant les connexions réseau, le score global du modèle s'est amélioré. Leur approche a atteint la deuxième place parmi environ 900 participants au concours de profilage des utilisateurs SMP 2017. Ce travail a montré que l'utilisation de graphes et la prise en compte de l'interactivité de l'utilisateur est une amélioration vers le regroupement des individus ; ainsi, détecter les communautés anormales. Une limitation majeure de ce travail est que les techniques utilisées ne peuvent pas être utilisées pour prédire la catégorie des futurs changements de comportement.

4.3 Critères d'évaluation

Deux études bibliographiques intéressantes ont traité en profondeur le même problème que notre travail, la détection d'anomalies dans le domaine des réseaux sociaux [123][124].

1. Dans [123], les auteurs ont discuté des techniques de Deep Learning (DL) dans les réseaux sociaux à différentes fins : analyse du comportement des utilisateurs, analyse commerciale, analyse des sentiments et détection des anomalies. Ils se sont attachés de près, dans la partie techniques de détection d'anomalies à base de DL, à commenter les travaux de la littérature sur la base de leur appartenance à l'apprentissage supervisé ou à l'apprentissage non supervisé.

2. Dans [124], les auteurs ont étudié la détection d'anomalies basée sur le DL dans divers domaines d'application : détection d'intrusion, détection de fraude, détection de logiciels malveillants, détection d'anomalie médicale, détection d'anomalie de journal, détection d'anomalie dans les médias sociaux, Internet of Things (IoT), Big Data Anomaly Detection, détection d'anomalies dans l'industrie, etc., et ils ont classé les travaux dans le domaine de la détection d'anomalies dans les réseaux sociaux par type de technique utilisée. Ils ont donc constaté que la nature hétérogène et dynamique des données présente des défis importants pour les techniques de DL.

3. Quant à notre travail, nous avons classé les travaux de la littérature par type de méthode (méthodes basées sur les activités, méthodes basées sur des graphes ou méthodes hybrides), où nous avons identifié des défis entre les méthodes qui utilisent le DL pour détecter des anomalies dans les réseaux sociaux. Ces défis apparaissent dans la capacité d'une méthode d'être capable d'analyser des données multimodales sur une structure multidimensionnelle, communautaire et dynamique. Encore plus, ces défis sont considérés comme les critères d'évaluation des méthodes impliquées. Sur la base de cette évaluation, notre travail peut être un complément aux deux enquêtes mentionnées ci-dessus.

4.3.1 Structure multidimensionnelle

Les propriétaires de comptes de réseaux sociaux perdent énormément de temps chaque jour dans des tâches chronophages inutiles et redondantes. Si nous additionnons toutes ces heures perdues en un mois, les résultats sont effrayants. Pour gagner du temps libre tout en assurant une présence constante sur les réseaux sociaux, la synchronisation des comptes est une solution très efficace. Cette synchronisation des contacts, des publications, etc. permet une meilleure maîtrise des informations et un croisement de toutes les données sur les plateformes concernées. Par exemple, si un utilisateur de réseaux sociaux possède un compte Facebook et Instagram, il peut les connecter, afin de disposer d'une passerelle. Cela facilite le partage de ses photos et vidéos entre les deux réseaux sociaux en un clic. C'est un excellent moyen de gagner du temps. La synchronisation des comptes peut être représentée par un graphe multidimensionnel où chaque dimension du graphe présente un réseau social. Entre deux nœuds du graphe, plusieurs liens peuvent avoir lieu. La figure 4.1 montre un réseau multidimensionnel extrait du site ResearchGate [12] où les nœuds représentent les auteurs de publications scientifiques et les liens représentent les différentes relations entre ces auteurs. Dans ce réseau, les nœuds s'interconnectent via cinq dimensions, à savoir : une connexion existe entre deux auteurs si :

— Dimension 1 : ils ont écrit au moins un article ensemble. (Liens bleus)
— Dimension 2 : l'un cite l'autre. (Liens rouges)
— Dimension 3 : ils ont la même affiliation. (Liens verts)
— Dimension 4 : ils travaillent sur le même projet. (Liens jaunes)
— Dimension 5 : ils ont plus de 10 compétences et expertises en commun. (Liens roses)

12. Discover scientific knowledge and stay connected to the world of science : https ://www.researchgate.net/.

Les avantages majeurs de l'analyse d'un réseau multidimensionnel apparaissent dans la collecte de diverses données sur les individus présents dans le réseau et l'explication du type de relations qui apparaissent entre ces individus.

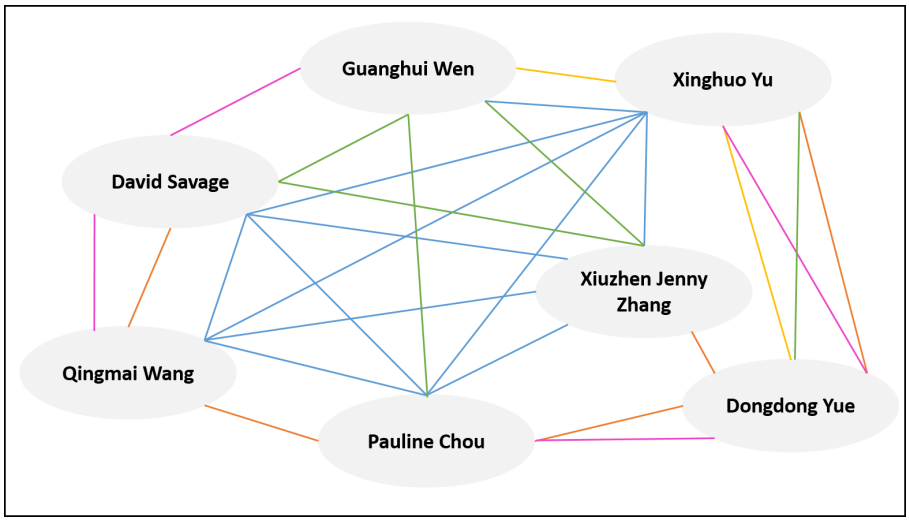

Figure 4.1: Exemple de connexion d'un réseau multidimensionnel.

4.3.2 Données multimodales

Le développement rapide des technologies de l'information a conduit à une évolution exponentielle de la masse de données hétérogènes enregistrées à partir de plusieurs méthodes d'acquisition. Ces données, multimodales et hétérogènes, se présentent souvent sous des formats différents (signaux, images, textes, vidéos, etc.), provenant de plusieurs sources (bases de données, réseaux sociaux, etc.) ce qui rend l'extraction de connaissances compliquée. Sur les réseaux, quatre grandes catégories de données sont présentes, à savoir :

— Les données de contenu textuel, qui incluent les publications, les commentaires, les légendes d'images, le texte dans une image, etc.,

— Les données de contenu d'image, telles que les photos publiées, les images de profil, etc.

— Les données de contenu vidéo, telles que les vidéos de profil, les vidéos publiées, etc.

— Les données à contenu numérique, telles que l'âge, le nombre d'amis, le nombre moyen de messages par jour, le temps passé à naviguer, etc.

4.3.3 Structure communautaire

Les utilisateurs des médias sociaux ont des liens particuliers car ils ont des affinités particulières, ou ont des caractéristiques similaires, ou partagent des intérêts, etc. Par conséquent, ils forment des communautés en fonction de leur similitude de comportement sur les réseaux sociaux [125]. Graphiquement, une communauté est constituée d'un ensemble de nœuds qui sont fortement liés entre eux, et faiblement liés aux nœuds situés à l'extérieur de la communauté. La valeur des informations extraites de la structure des communautés est multiple : identifier des profils types, mener des actions ciblées, mieux ajuster les recommandations, identifier les acteurs principaux/influents, analyser le comportement des individus, déterminer les comportements anormaux, etc. Plusieurs études ont montré l'apport de l'extraction communautaire dans le domaine de la détection des comportements anormaux dans les réseaux sociaux [22][24][9][23][26][110]. La figure 4.2 présente un exemple de partitionnement d'un graphe en quatre sous-graphes denses (communautés) faiblement liés les uns aux autres.

4.3.4 Comportement dynamique

Avec les progrès des technologies de l'information, les réseaux sociaux sont devenus un moyen de communication libre mais aussi un nouveau moyen pour les entreprises ou les politiques d'influencer les décisions des internautes. Cette influence est la cause majeure du changement de comportement d'un individu au

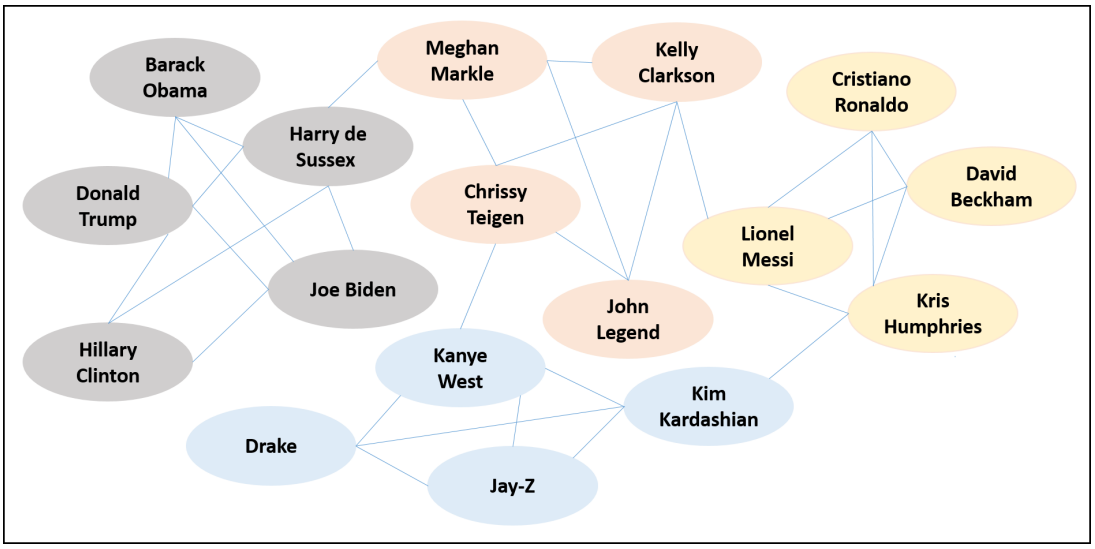

Figure 4.2: Exemple de construction de communautés.

cours du temps. Ainsi, les réseaux sociaux admettent une structure sociale dynamique où ils désignent spécifiquement l'objet d'étude défini par la structure des relations interpersonnelles établies entre les personnes autour de certains contenus relationnels. Ils évoluent, se reconfigurent, se déploient ou se contractent en permanence. Leurs déplacements sont liés à des changements de relations, à l'apparition de nouveaux événements, à l'émergence de communications, etc. Ces réseaux humains de communication sont modélisés avec des métriques liées à l'ampleur du changement d'état de leurs nœuds. La prise en compte de l'aspect dynamique dans l'analyse des réseaux sociaux pour détecter les comportements anormaux est une nécessité. Plusieurs études sont parvenues à remettre en cause ce besoin primordial [98][99][100][36][110][121].

4.4 Discussion

Chacune des méthodes de détection des comportements anormaux dans les réseaux sociaux présentées ci-dessus a ses avantages et ses inconvénients. De ce fait,

il est essentiel de comprendre quelle est la meilleure méthode dans ce contexte ou quelle méthode doit être développée pour satisfaire tous les besoins d'une détection d'anomalie efficace.

Dans la section précedente, suite à l'analyse d'un grand nombre de réseaux sociaux, nous avons proposé quatre critères d'évaluation, qui présentent des mesures de performance pour une méthode de détection d'anomalies. Ainsi, malgré les résultats probants des méthodes existantes, aucune d'entre elles n'a réussi à répondre à tous ces critères. Le tableau 4.1 présente les forces et les faiblesses liées aux différentes méthodes critiquées. Compte tenu de l'importance de la structure multidimensionnelle des réseaux sociaux, les méthodes structurelles illustrées dans la section 4.2.2 semblent être le meilleur choix. Cependant, elles ne répondent pas au critère d'analyse de données multimodales par le fait que cette catégorie de méthodes n'aborde pas le comportement des utilisateurs en amont c'est-à-dire qu'elle ne calcule que des métriques graphiques pour connaître le degré d'implication de chaque utilisateur par rapport aux autres. Une analyse approfondie des différents types de données des activités doit être incluse. Par conséquent, une méthode hybride est sans aucun doute plus appropriée. Les méthodes hybrides présentées dans la section 4.2.3 montrent encore des faiblesses dans l'analyse d'une structure multidimensionnelle. Par conséquent, nous optons pour une méthode hybride traitant un réseau multidimensionnel comme une opportunité d'améliorer la valeur des solutions actuelles. Enfin, nous affirmons qu'à notre connaissance, il n'existe pas d'étude existante dans la littérature, qui remette en cause l'ensemble des besoins évoqués ci-dessus. Dans l'ensemble, les résultats obtenus jusqu'à présent sont importants et peuvent être considérés comme les fondations sur lesquelles nous pouvons construire de nouvelles recherches pour atteindre de nouvelles réalisations ambitieuses.

Table 4.1: Évaluation des approches-clés

Approche	Type	SM	DM	DC	CD
[37]	*comportemenal*	×	✓	×	×
[36]	*comportemenal*	×	✓	×	✓
[98]	*comportemenal*	×	×	×	✓
[100]	*comportemenal*	×	×	×	✓
[28]	*comportemenal*	×	✓	×	×
[29]	*comportemenal*	×	×	×	×
[106]	*comportemenal*	×	×	×	×
[107]	*comportemenal*	×	×	×	×
[89]	*comportemenal*	×	×	×	✓
[24]	*structurel*	×	×	×	×
[9]	*structurel*	×	×	✓	×
[25]	*structurel*	×	×	×	×
[22]	*structurel*	×	×	✓	×
[23]	*structurel*	×	×	×	✓
[26]	*structurel*	×	×	✓	✓
[27]	*structurel*	✓	×	×	×
[31]	*structurel*	×	×	×	×
[118]	*structurel*	×	×	✓	✓
[119]	*structurel*	×	×	✓	×
[110]	*structurel*	×	×	✓	✓
[109]	*structurel*	×	×	×	×
[120]	*structurel*	✓	×	×	✓
[121]	*hybride*	×	×	×	✓
[35]	*hybride*	×	✓	×	×
$Notre framework$	*hybride*	✓	✓	✓	✓

4.5 Conclusion

Dans ce chapitre, un état de l'art sur la détection des comportements anormaux dans les réseaux sociaux a été présenté. Nous avons catégorisé les méthodes existantes en trois familles : méthodes comportementales, méthodes structurelles et méthodes hybrides. Ces méthodes présentent plusieurs inconvénients en dépit

de leur efficacité. Cette analyse des travaux existants a fait l'objet d'une publication dans le journal "Multimedia Systems (2021)" [13]. Dans le chapitre suivant, nous aborderons notre première contribution qui s'intéresse à l'exploration d'une structure monodimensionnelle en considérant une analyse comportementale manipulant des données textuelles et images.

13. https ://link.springer.com/article/10.1007/s00530-020-00731-z

Chapitre 5

Modèle de détection et de prédiction des comportements anormaux sur Twitter

Sommaire

5.1 Introduction

Dans le monde numérique d'aujourd'hui, la lutte contre le terrorisme est considérée comme l'une des plus hautes priorités des départements de la défense du monde entier. Les chercheurs et les gouvernements investissent dans la création de technologies de l'information avancées pour identifier et lutter contre le terrorisme grâce à une analyse à grande échelle des données en ligne, en particulier des réseaux sociaux. Des groupes terroristes militants exploitent ces réseaux dans le but de promouvoir leurs organisations et de recruter les personnes les plus naïves au sein de leurs communautés dangereuses ; le réseau social le plus utilisé par ces groupes est Twitter. Cependant, les approches existantes sont peu efficaces ou n'étudient pas les comportements de ces utilisateurs malveillants et ne reposent que sur des données textuelles fournies par les utilisateurs. Dans ce chapitre, nous proposons un nouveau modèle de détection et de prédiction de l'influence des comportements des terroristes sur les réseaux sociaux en analysant à la fois leur contenu texte et image publié en faisant recours à diverses techniques d'apprentissage automatique. Nous suggérons ainsi la construction de graphe social terroriste ou sociogramme qui est un moyen utile pour toute analyse ultérieure des réseaux sociaux. Cette visualisation graphique présente une information pertinente et parvient à une meilleure compréhension des comportements tout en découvrant des informations souvent

implicites et précieuses.

5.2 Architecture générale de notre modèle

La motivation majeure sur laquelle s'appuie notre travail, c'est le manque de travaux en rapport avec la compréhension des réseaux terroristes. En effet, les travaux antérieurs se concentrent soit sur le problème de classification binaire qui détermine si les utilisateurs sont pro-ISIS ou anti-ISIS, soit sur le problème de l'analyse des réseaux sociaux où différentes approches tendent à analyser les réseaux avec différentes techniques sans exploiter les données textuelles ou les images. Par conséquent, ces approches sont principalement manuelles. L'objectif de notre modèle est d'automatiser le problème de classification et d'analyse de réseau et d'aboutir à une modélisation des résultats compréhensible sous forme de sociogramme. En ce qui concerne la classification, nous utiliserons deux sous-modèles : (1) un sous-modèle de classification de texte où les données sont constituées de publications d'utilisateurs, de commentaires, etc. et (2) un sous-modèle de classification d'images où les données sont constituées d'images de profil, de publications avec images, d'images commentaires, etc. De plus, nous prédirons les utilisateurs terroristes potentiels, ce sont les utilisateurs qui ont été détectés comme anti-ISIS mais qui appartiennent toujours à un réseau terroriste. Le sous-modèle de prédiction sera basé sur la technique de Collaborative Filtering (CF) utilisée par les systèmes de recommandation modernes [126]. Notre résultat final est un graphe représentant le réseau terroriste où les nœuds représentent soit des pro-ISIS, soit des anti-ISIS détectés et prédits.

Le workflow illustré à la figure 5.1 donne un aperçu de notre modèle proposé. Ce dernier se compose de trois phases principales : (1) détection des nœuds terroristes : chaque noeud à l'intérieur du réseau sera analysé par nos sous-modèles de texte et

image, (2) prédiction des nœuds terroristes : chaque nœud classé comme négatif (anti-ISIS) sera analysé par notre sous-modèle de prédiction, et (3) remplissage du graphe de réseau terroriste : chaque fois qu'une analyse de détection ou de prédiction a lieu, le graphe construit se met à jour en fonction des résultats de nos sous-modèles.

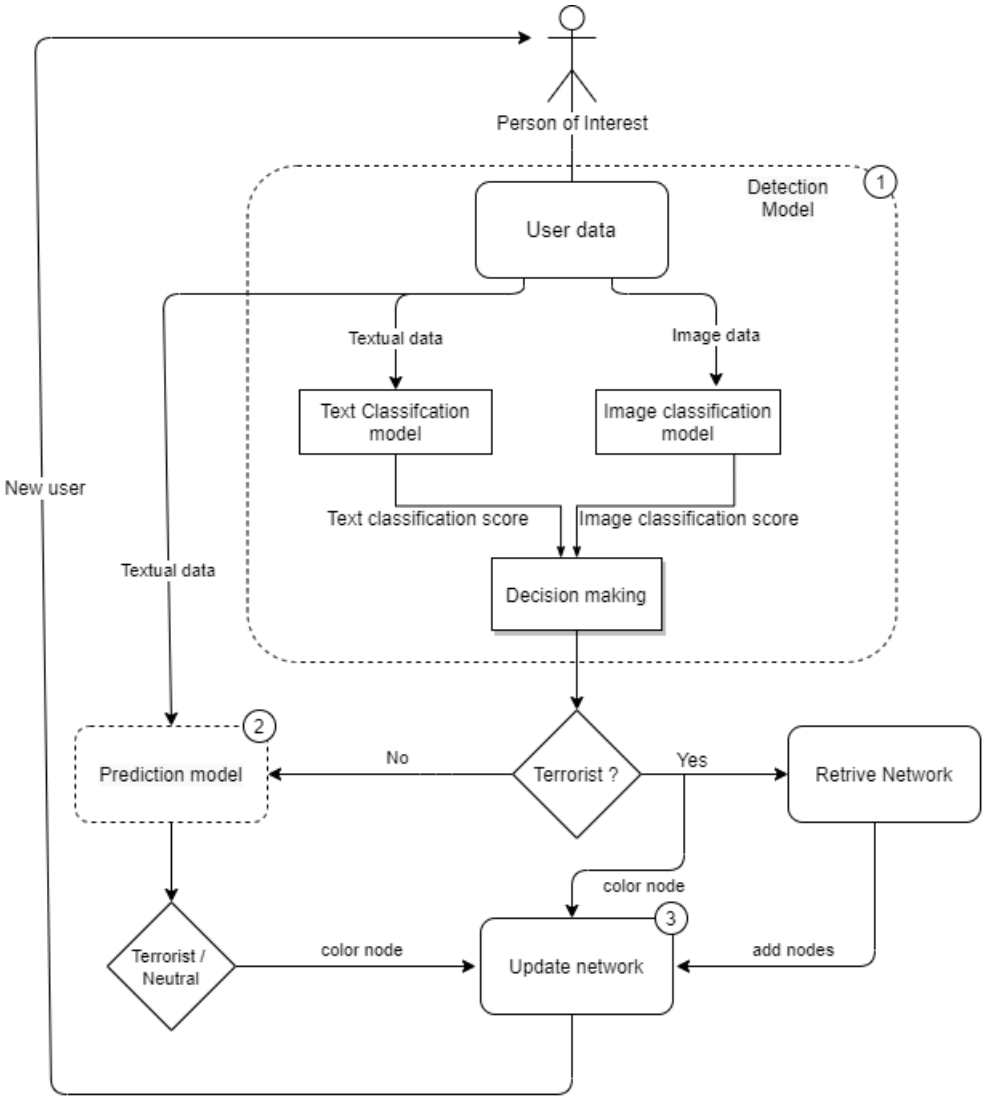

Figure 5.1: Workflow du modèle proposé.

5.3 Description de la méthodologie de notre modèle

5.3.1 Sous-modèle de détection

Le sous-modèle de détection est composé de deux sous-modèles de classifications, un pour les données textuelles et un pour les données images et un module d'aide à la décision en charge de l'agrégation de deux scores générés automatiquement des deux sous-modèles. Dans ce qui suit, nous donnons un aperçu des sous-modèles respectifs.

5.3.1.1 Sous-modèle de classification de texte

Le processus de classification des données d'entrée textuelles se compose de trois parties principales, comme illustré dans la figure 5.2. (1) Application du NLP : le prétraitement des données textuelles afin de disposer de données structurées prêtes à l'emploi et faciles à comprendre et traiter. Le processus d'analyse des données textuelles s'effectue en quatre étapes majeures : le marquage, l'annotation, la coréférence et l'analyse des sentiments [127]. (2) Word Embedding (WE) [128] : l'utilisation du modèle N-gramme qui estime la probabilité du dernier mot par rapport aux mots précédents. Ce choix a été motivé par le fait que les n-grammes vont être réutilisés et très utiles pour notre sous-modèle de prédiction [129]. (3) Classification : le contenu textuel est maintenant sous une forme numérique, compréhensible par la machine et prêt à être utilisé par n'importe quel classificateur de machine learning. L'exécution du processus de classification de texte permet de générer un score de profil basé sur les données textuelles.

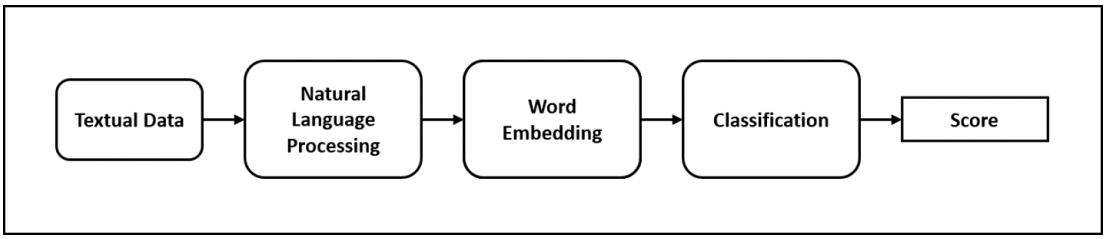

Figure 5.2: Conception du modèle de classification de texte.

5.3.1.2 Sous-modèle de classification d'images

Les CNNs sont connus pour être les meilleurs classificateurs lorsqu'il s'agit du traitement des données d'image [82]. Afin d'aboutir à notre objectif, nous commençons par fournir l'image d'entrée à la couche de convolution, nous prenons la convolution avec les filtres sélectionnés, appliquons la couche pooling pour réduire les dimensions, puis nous ajoutons ces couches plusieurs fois, nous aplatissons la sortie et alimentons une couche entièrement connectée afin d'être prêt à entraîner le modèle. Nous avons également utilisé certaines fonctionnalités courantes de la technique Data Augmentation (mise en miroir, recadrage aléatoire, rotation, ajustement de l'échelle, changement de couleur, etc.). Cela permet d'élargir l'ensemble de données et d'augmenter la quantité de données pertinentes en augmentant la diversité des données d'entraînement, ce qui est essentiel pour développer un modèle robuste et améliorer les performances des modèles d'apprentissage profonds. Comme le sous-modèle de classification de texte, un score spécifique au sous-modèle de classification d'images est généré.

5.3.1.3 Composant de prise de décision

Ce module calcule le score final en fonction des scores fournis par les sous-modèles de classification de textes et d'images. L'équation 5.1 calcule le score final

pour l'utilisateur S_u ayant les scores s_1, s_2 et les poids α_1, α_2 attribués respectivement aux deux modèles de classification. Ces poids sont calculés en tenant compte des métriques des classificateurs de texte et d'image ainsi que de la taille des ensembles de données collectées.

$$S_u = \sum_{i=1}^{2} \alpha_i . s_{u_i} \tag{5.1}$$

Sur la base de ce score final, une décision est prise pour savoir si un utilisateur est pro-ISIS (terroriste) ou anti-ISIS (non terroriste) en définissant un seuil γ. Si le score est supérieur ou égal à ce seuil, une étiquette positive est choisie sinon, une étiquette négative est attribuée comme indiqué dans l'équation 5.2.

$$\begin{cases} S_u >= \gamma \Rightarrow Positive(Pro - ISIS) \\ S_u < \gamma \Rightarrow Négative(Anti - ISIS) \end{cases} \tag{5.2}$$

5.3.2 Sous-modèle de prédiction

Ce sous-modèle utilise la technique de Collaborative Filtering basée sur les préférences similaires des utilisateurs. Ce choix est motivé par le fait que les personnes obtiennent souvent les meilleures recommandations de quelqu'un ayant des goûts similaires. Dans notre cas, ces recommandations sont les n-grammes extraits de notre sous-modèle précédent et les notes sont les valeurs de ces n-grammes (calculées en fonction de leur fréquence) dans les publications créées et partagées par les utilisateurs. Grâce à cette technique, le sous-modèle de prédiction prévoit l'influence potentielle future du terrorisme pour un utilisateur qualifié de négatif (anti-ISIS). Nous sommes ainsi en mesure de prédire quels utilisateurs sont susceptibles d'être attirés par des groupes terroristes.

5.3.3 Sociogramme du réseau terroriste

La dernière étape de notre modèle consiste à la construction d'un sociogramme terroriste [1]. Cette représentation est très utile à des fins de visualisation, elle rend les données de sortie plus compréhensibles pour les non-experts dans les domaines de l'analyse des réseaux sociaux et de l'apprentissage automatique. C'est l'étape la plus importante de notre travail même si c'est la moins compliquée parmi toutes les étapes précédentes.

Afin de construire le graphe du réseau terroriste, nous devons choisir le type et la forme du graphe. Puisqu'il s'agit de données complexes, il est préférable d'utiliser des graphes unidirectionnels. Ces graphes comprennent :

— Des sommets qui représentent les utilisateurs du réseau social

— Des arêtes qui représentent le lien d'amitié entre deux nœuds. Le lien d'amitié sera calculé sur le nombre d'interactions publiques entre deux nœuds, les nœuds avec le plus d'interactions seront ajoutés au graphe.

Quant à la forme du graphe, nous avons choisi d'utiliser des graphes en étoile qui sont parfaits pour représenter le réseau de chaque utilisateur individuel. Les graphes en étoile sont des graphes biparties complets avec un nœud interne et k feuilles, donc un graphe en étoile noté S_k comprend $k + 1$ sommets et k arêtes (Figure 5.3).

5.4 Résultats et interprétation

5.4.1 Collecte de données

Pour entrainer notre classificateur, nous avons besoin d'étiquettes positives et négatives afin d'aider le classificateur à distinguer le contenu des médias sociaux

1. Sociogram : https ://www.merriam-webster.com/dictionary/sociogram.

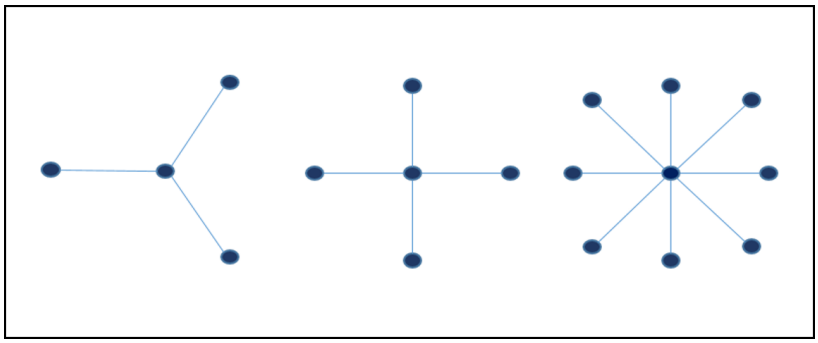

Figure 5.3: Graphes en étoile S_3, S_4 et S_8

pro-ISIS du contenu d'actualités générales.

5.4.1.1 Données hors ligne

Données textuelles Les modèles de détection et de prédiction reposent sur des données textuelles, c'est pourquoi il est très important de choisir les ensembles de données les plus riches.

— **Etiquettes positives :** Nous sommes tombés sur une grande quantité d'ensembles de données qui ne correspondent pas à nos besoins. Nous nous sommes intéressés aux ensembles de données précédemment partagés entre les utilisateurs pro-ISIS dans les réseaux sociaux. Nous avons découvert le "How ISIS uses Twitter" [2] qui contient 17350 tweets de plus de 110 comptes pro-ISIS. Il comprend les attributs suivants : Nom, Nom d'utilisateur, Description, Emplacement, Nombre d'abonnés au moment où le tweet a été téléchargé, Nombre de statuts de l'utilisateur lorsque le tweet a été téléchargé, Date et horodatage du tweet et du tweet lui-même. Certains des tweets de l'ensemble de données sont écrits en arabe, nous avons donc utilisé un outil de traduction publique, l'API Google Translate, chaque fois que nous détectons des caractères arabes.

2. How ISIS Uses Twitter : https ://www.kaggle.com/fifthtribe/how-isis-uses-twitter.

— **Etiquettes négatives :** Nous avons utilisé la base de données GTD [113] dans le cadre de notre ensemble de données étiquetées négatives. L'ensemble de données contient des informations sur plus de 180000 attaques terroristes dans le monde entier depuis 1970. Nous avons filtré l'ensemble de données sur les événements survenus depuis 2002 et extrait le contenu de la colonne "summary" qui contient des résumés de chacune des attaques. Un exemple des données que nous utilisons est illustré à la figure 5.4.

Notre ensemble de données final contient 17350 échantillons positifs et 26894 négatifs. Le tableau 5.1 montre la quantité d'échantillons de données textuelles utilisées pour l'apprentissage.

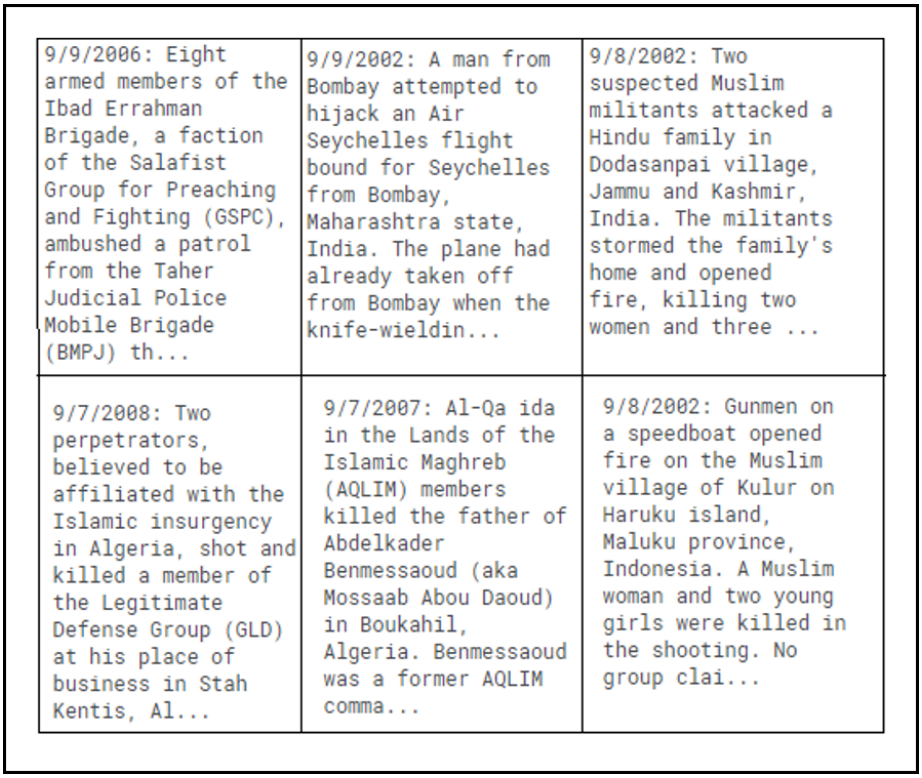

Figure 5.4: Extrait de GTD de la colonne récapitulative

Données images

— **Etiquettes positives (Pro-ISIS) :** Pour le contenu de l'image, nous avons collecté les données des précédents comptes pro-ISIS bannis sur Twitter. Nous avons pu extraire les photos de profil, malheureusement, le contenu multimédia des publications n'était pas disponible. Les données extraites comprennent 60 images. Nous avons également extrait des images de l'API de recherche Google, montrant du contenu pro-ISIS et nous avons filtré manuellement les échantillons et les avons ajoutés à l'ensemble de données des échantillons positifs.

— **Etiquettes négatives (Anti-ISIS) :** Quant à notre ensemble de données étiquetées négatives, nous avons également extrait des données de l'API de recherche Google contenant des images de personnes d'apparence neutre et des images d'articles de presse ayant les mots-clés ISIS, DAESH, TERRORIST, etc. Le nombre d'échantillons utilisés pour l'apprentissage et le test du classificateur est indiqué dans le tableau 5.1.

Table 5.1: Ensembles de données extraits de contenu textuel et d'image

Ensemble de données	Etiquettes positives	Etiquettes négatives	Total
Textuel	17350	26894	44244
Image	341	308	649

5.4.1.2 Données en ligne

Comme nous l'avons mentionné précédemment, nous ne pouvons accéder qu'au point de terminaison de l'API des médias sociaux Twitter grâce à l'accessibilité et à la diversité de l'extraction de données qu'il fournit, contrairement à d'autres plateformes de médias sociaux qui autorisent uniquement à extraire les données du propre utilisateur uniquement [3]. Si ces données sont nombreuses, elles se déclinent sous deux formes :

3. Twitter for Developers. https ://developer.twitter.com.

— **Tweet Metadata :** ils contiennent des informations utiles sur chaque tweet. Le point de terminaison de l'API peut offrir de nombreuses données telles que le nombre de favoris (favoriteCount), le nombre de retweets (retweetCount) ou la date à laquelle le tweet a été posté. L'information la plus importante est le "inReplyToScreenName", qui indique si un message est une réponse à un autre tweet ou est un message original. C'est le champ le plus important de notre analyse, car il est utile pour récupérer la liste des interactions publiques directes entre deux utilisateurs.

— **User Metadata :** ils contiennent des informations générales sur chaque utilisateur telles que la description du profil, les abonnés et les abonnements et le nombre de tweets qu'ils ont publiés.

5.4.2 Résultats du sous-modèle de détection

5.4.2.1 Résultats du sous-modèle de classification de texte

Dans cette section, nous utilisons différents classifieurs (NB, SVM et LR) pour notre modèle de classification de texte afin de choisir le meilleur. L'implémentation de ces classifieurs est incluse dans la bibliothèque Scikit-learn [130]. Nous avons convenu d'utiliser 70% des échantillons pour l'entrainement du modèle et 30% pour les tests. Chaque algorithme d'entraînement et de test a été implémenté sur la même machine et dans les mêmes conditions séparément. Les résultats de la classification sont présentés dans le tableau 5.2.

Table 5.2: Résultats de la classification pour le sous-modèle de texte

Classifieur	Précision	Rappel	F-score	Temps d'entrainement
Support Vector Machine	0.907	0.902	0.904	6h 48min 33s
Naive Bayes	**0.904**	**0.899**	**0.900**	**1min 11s**
Logistic Regression	0.899	0.854	0.875	39.9s

Nous avons choisi de travailler avec le classifieur NB malgré le fait que le F-score du classifieur SVM était légèrement plus élevé. En effet, le SVM prend plus de

temps à s'entraîner pour une différence minimale dans les résultats de classification. Comme nous devons souvent réentraîner le modèle, un temps d'entraînement réduit est requis.

5.4.2.2 Résultats du sous-modèle de classification d'image

Le CNN est la norme pour la reconnaissance d'images [82]. Nous avons utilisé Keras comme un framework qui fonctionne sur TensorFlow et qui permet une expérimentation rapide grâce à une API de haut niveau, conviviale et extensible [4]. L'architecture de notre modèle est composée de trois couches, avec respectivement 16 neurones, 8 neurones et 1 neurone. Les deux premières couches ont une activation "relu" car ses performances sont prouvées, et la dernière couche a une activation "sigmoïde" car nous traitons un problème de classification binaire et car c'est notre couche de sortie. Le modèle est compilé avec "binary_crossentropy" comme fonction de perte et "adam" comme optimiseur.

Nous avons utilisé un CNN en combinaison avec deux techniques d'optimisation, à savoir : TL et DA. Tout d'abord, nous avons implémenté les fonctions DA, puis nous avons défini les connaissances acquises du modèle de base à utiliser. Dans le tableau 5.3, nous présentons les différents scores de combinaison en utilisant nos couches CNN, avec et sans le modèle pré-entraîné et avec et sans les données générées à partir de DA. Bien que les scores soient mesurés par les mêmes données de test, les données d'entraînement diffèrent lors de l'utilisation du DA. L'utilisation conjointe de DA et de TL a permis d'obtenir le meilleur score et un temps d'entraînement réduit. Par conséquent, nous avons utilisé cette combinaison dans notre modèle global.

4. Keras. https ://keras.io.

Table 5.3: Résultats de la classification pour le sous-modèle d'image

Classifieur	Précision	Rappel	F-score	Temps d'entrainement
CNN	0.761	0.713	0.735	4min 20s
CNN + DA	0.861	0.849	0.853	4min 46s
CNN + TL	0.874	0.867	0.869	8min 34s
CNN + DA + TL	**0.929**	**0.913**	**0.920**	**9min 36s**

5.4.2.3 Composant de prise de décision

Le composant de prise de décision est chargé de l'exécution des deux tâches suivantes :

— Tâche 1 : L'interprétation d'un score de sortie du modèle global en fonction des poids des sous-modèles (Équation 5.1). Pour calculer un poids pour chaque sous-modèle, nous nous sommes intéressés à extraire des informations pertinentes afin de représenter les données par plusieurs caractéristiques. Ces caractéristiques sont présentées par les types de données utilisées lors de l'apprentissage du modèle global, à savoir : les textes et les images. L'essentiel à ce stade est de calculer l'importance de l'impact de chacune de ces caractéristiques sur le modèle entraîné. Nous avons donc utilisé le modèle XGBoost [5] pour lequel la bibliothèque Python XGBoost propose plusieurs calculs. Le principe du modèle adopté repose sur l'occurrence de la caractéristique dans le modèle entraîné. La figure 5.5 présente les occurrences de chaque caractéristique où il est très visible que les données textuelles ont une légère importance dans l'apprentissage du modèle par rapport aux données images. La figure 5.6 montre les valeurs des poids calculés par le modèle XGBoost.

— Tâche 2 : L'estimation de la valeur du seuil pour la classification des différents scores obtenus. Quant au seuil de prise de décision, nous l'avons fixé à partir de l'utilisation de la courbe rappel-précision pour différentes valeurs

5. https ://github.com/aureliale/features_importance

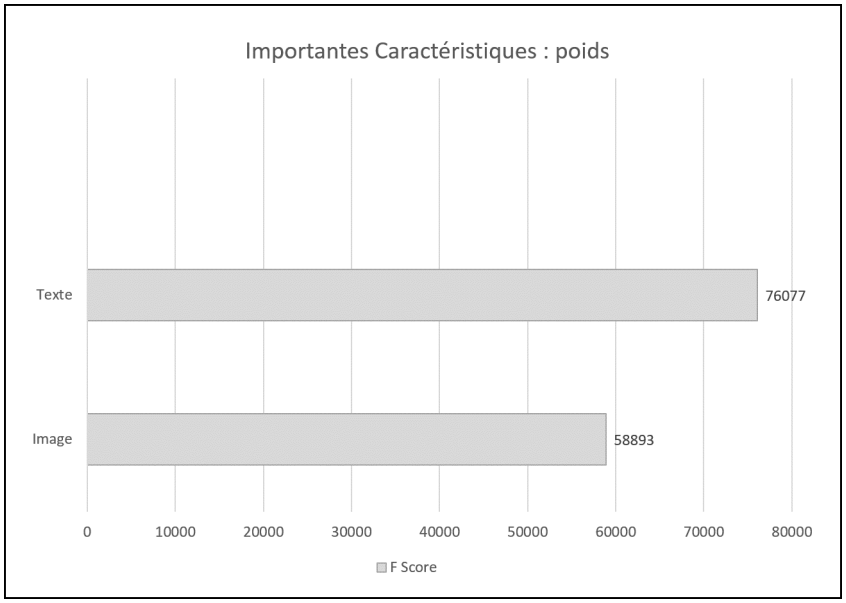

Figure 5.5: Occurences des caractéristiques.

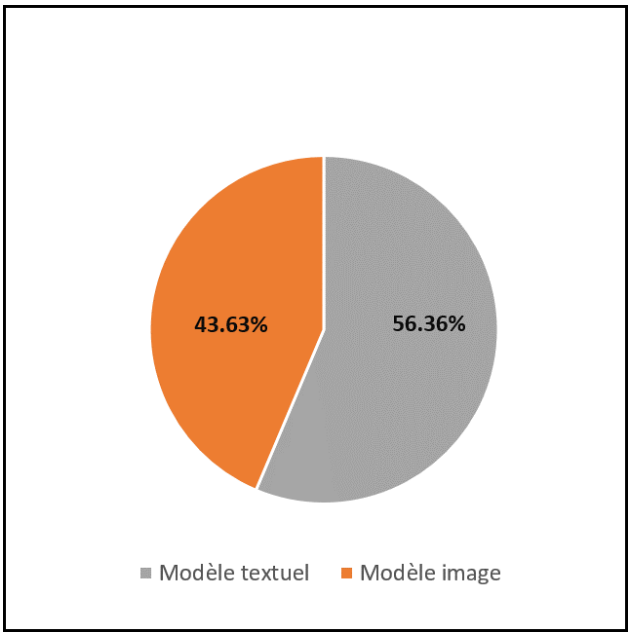

Figure 5.6: Valeurs des poids estimés.

de seuil. Nous avons sélectionné la meilleure valeur pour le seuil de décision afin qu'il donne une précision élevée ou un rappel élevé. Les valeurs de score de décision supérieures à ce seuil sont considérées comme une classe pro-ISIS. Par conséquent, nous avons choisi la valeur du seuil de décision qui augmenterait la précision mais ne diminuerait pas considérablement le rappel. La valeur adéquate est d'environ 0,62. Cette valeur de seuil est calculée automatiquement à chaque changement sur les scores de sortie du modèle global.

En général, les sous-modèles entraînés dépendent des données, ce qui rend nécessaire de les réentraîner lorsque les données changent afin d'obtenir une prédiction précise de l'étiquette. Pour résoudre ce problème, nous avons développé un module qui réentraîne et rétablit un modèle dès que les données changent. Chaque modèle est stocké avec son dernier score dans une base de données et une fonction python vérifie si le score est amélioré après le réentraînement du modèle sur les données d'un nouvel utilisateur terroriste.

5.4.3 Résultats du sous-modèle de prédiction

Pour implémenter le sous-modèle de prédiction, nous avons utilisé la librairie librec Python [6]. Initialement codé en Java, librec est une bibliothèque d'algorithmes de systèmes de recommandation. Elle comprend une variété d'implémentations pour les algorithmes de recommandation, le filtrage basé sur le contenu et le filtrage collaboratif [131]. Nous avons eu recours à l'utilisation de deux algorithmes. Les meilleurs résultats de classification ont été obtenus à partir de l'algorithme de Item KNN [131]. Ces résultats sont présentés dans le tableau 5.4.

6. Librec : Python library. https ://pypi.org/project/librec.

Table 5.4: Résultats du sous-modèle de prédiction.

	Non-negative matrix factorization (NMF)			**Item k-Nearest Neighbors (Item KNN)**		
	Précision	Rappel	F-score	**Précision**	**Rappel**	**F-score**
Pro-ISIS	0.742	0.613	0.670	**0.792**	**0.655**	**0.702**
Anti-ISIS	0.810	0.610	0.695	**0.860**	**0.660**	**0.746**

Les résultats de la classification n'étaient pas très optimaux, car nous observons que le F-Score pour les deux classes (Pro-ISIS et Anti-ISIS) est encore un peu faible. Nous essayons de les améliorer dans les contributions qui suivent (Chapitre 6 et chapitre 7).

5.4.4 Construction du sociogramme du réseau terroriste

La dernière étape consiste à construire un réseau pour chaque terroriste détecté. Ce réseau est mis à jour chaque fois qu'un nouvel événement de détection ou de prédiction se produit.

1. Si un nœud est détecté comme étant pro-ISIS : lorsqu'un nœud est détecté, premièrement, nous récupérons ses amis les plus proches dans le réseau (ceux qui interagissent le plus avec lui) en déclenchant d'abord la méthode "statuses" de l'API twittersearch. Deuxièmement, nous analysons tous les nœuds avec lesquels il interagit le plus publiquement en taguant et en retweetant leur contenu (@username, RTusername) car ce sont les seuls moyens d'identifier publiquement les interactions directes entre les utilisateurs sur Twitter. Les nœuds qui sont colorés en rouge sont ceux qui sont détectés comme pro-ISIS. Leurs amis les plus proches sont ajoutés au réseau qui leur est directement connecté.
2. Si un nœud est prédit comme un nœud potentiel pro-ISIS ou Anti-ISIS : Cette étape est très utile pour une validation plus poussée et une meilleure compréhension du réseau terroriste. Un nœud peut être soit prédit comme

un nœud potentiellement dangereux à l'intérieur du réseau et par conséquent qui doit être surveillé, soit prédit simplement un utilisateur Anti-ISIS et alors ne constitue pas une menace contrairement à d'autres nœuds dangereux.

Pour cela, nous avons utilisé la bibliothèque python NetworkX. Cette bibliothèque est très utile pour les travaux liés à la théorie des graphes et à l'analyse des réseaux sociaux [132]. Un exemple de notre implémentation de sortie finale est illustré dans la figure 5.7. Nous représentons les utilisateurs détectés et prédits dans chaque catégorie par trois couleurs distinctes :

- Rouge : représente les nœuds pro-ISIS détectés.
- Vert : représente les nœuds détectés et prédits comme étant anti-ISIS.
- Violet : Représente un potentiel anti-ISIS à recruter.

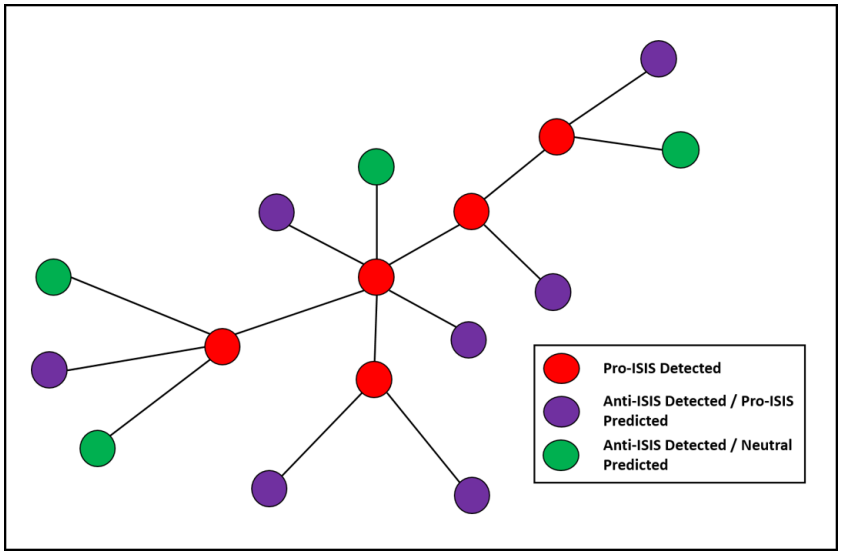

Figure 5.7: Sociogramme du réseau terroriste.

Avec ce sociogramme terroriste, toute une communauté terroriste présente sur les réseaux sociaux serait exploitable pour tout type d'analyse des réseaux sociaux

aux côtés de leurs données. Avec cela, nous pouvons valider les classes de détection et de prédiction de chaque nœud à l'intérieur du réseau. Nous pourrions également recycler nos modèles afin d'avoir de meilleurs scores de classification.

5.5 Conclusion

Dans ce chapitre, nous avons proposé un nouveau modèle de détection et de prédiction des comportements terroristes (pro-ISIS) sur Twitter. Notre modèle prend en charge différents types de données d'entrée telles que les textes et les images. Nous avons convenu que nous devrions inclure une classification des images puisque les images sont le type de contenu le plus influent dans les médias sociaux. Comme discuté dans les travaux connexes, la plupart des méthodes existantes ne reposent pas sur des données d'image. Cette contribution a été publiée dans les actes de la conférence IEEE SMC 2021 [7]. Malheureusement, nous n'avons pu utilisé que Twitter en raison de l'accessibilité qu'il offre pour le partage de contenu sur les réseaux sociaux. Comme prochaine contribution (Chapitre 7), nous avons étendu cette approche en explorant le comportement des utilisateurs dans plusieurs médias sociaux. L'idée vient du fait qu'une personne pourrait se comporter différemment dans différentes plateformes en ligne. Dans le chapitre suivant (Chapitre 6), nous traiterons l'aspect structurel multidimensionnel pour être combiné à son tour avec la contribution décrite dans le Chapitre 7.

7. https ://ieeexplore.ieee.org/document/9659253

Chapitre 6

Méthode de détection des comportements anormaux sur la base de l'analyse des relations dans une structure multidimensionnelle

Sommaire

6.1 Introduction

Suite aux meilleurs résultats de notre première contribution présentée dans le chapitre précédent, nous passons à aborder la problématique de la détection d'anomalies dans les réseaux d'informations monodimensionnels aux réseaux d'informations multidimensionnels afin de bien pouvoir prouver l'importance d'analyse de ces derniers. La portée de la détection d'anomalies ainsi que les travaux limités à ce propos dans les réseaux multidimensionnels ont motivé notre intérêt à élaborer une méthode qui détecte les nœuds atypiques. Ce manque d'approches qui se rattachent à la détection d'anomalies dans le contexte des réseaux multidimensionnels est possiblement lié, d'une part, à la structure topologique non triviale des réseaux multidimensionnels, et d'autre part, à l'absence d'une définition formelle du concept d'anomalies dans les réseaux multidimensionnels. Dans ce chapitre, nous traitons l'anomalie comme un comportement qui a une déviation par rapport aux autres mais pas obligatoirement éloigné des autres.

6.2 Méthode proposée

De nos jours, les réseaux sociaux se multiplient et sont parfois difficiles à gérer. Le même utilisateur peut avoir plusieurs comptes sur différents réseaux sociaux. Cependant, la synchronisation entre ces différents comptes est nécessaire.

Par exemple, la synchronisation permet à l'utilisateur de publier une photo simultanément via Instagram et Facebook en un clic. A notre connaissance, seuls les travaux [27][120] ont été effectués dans le domaine de la détection des anomalies dans des réseaux multidimensionnels.

6.2.1 Notation

Dans notre méthode, nous nous inspirons de la notation utilisée dans [10] pour modéliser la structure d'un graphe multidimensionnel. Un graphe multiple non orienté G est défini par le triplet (V, E, D) où V est l'ensemble des nœuds, E est l'ensemble d'arêtes et D est l'ensemble de dimensions. Un arc $e \in E$ est un triplet (u, v, d) où $u, v \in V$ et $d \in D = \{\text{Twitter, Facebook, Instagram, ... }\}$. Le triplet (u, v, d) spécifie que les nœuds u et v sont connectés par un arc e qui appartient à la dimension d.

Les propriétés locales du graphe social doivent être déterminées afin de nous aider à détecter les nœuds atypiques. Ces propriétés désignent un seul nœud par *un ego* et son voisinage à un premier niveau par *un egonet*. Notre méthode fonctionne en trois phases : (1) la détection des communautés dans les différentes dimensions du graphe, (2) l'estimation d'un score total d'anomalie pour chaque nœud sur toutes les dimensions et (3) la classification automatique des scores estimés pour identifier le type du comportement de chaque noeud.

6.2.2 Détection des communautés dans les différentes dimensions

La détection de communautés a pour objectif de regrouper les nœuds du graphe en groupes partageant des caractéristiques communes. Cette disposition est vraie dans le contexte des réseaux sociaux en ligne [125]. Les utilisateurs des réseaux

sociaux se comportent de manière à former des communautés en fonction de leurs préférences et de leurs intérêts communs. Diverses techniques ont été présentées dans divers travaux pour résoudre ce problème général. Un travail intéressant [9] a montré que la contribution de détection de communautés apparaît dans l'utilité des informations extraites de la structure des communautés formées. Ces informations facilitent l'analyse du comportement d'un utilisateur et permettent l'identification d'un comportement anormal.

Dans [24], les auteurs ont défini qu'un $egonet(u)$ forme une communauté avec un $egonet(v)$ si au moins la moitié des nœuds du plus petit egonet se connectent à l'autre egonet. L'application de l'équation 6.1 [9] permet de calculer les communautés d'un graphe.

$$Com(u,v) \quad = \quad \begin{cases} 1, \quad if \quad degree(u,v)_{norm} \geq \min\left(|u|,|v|\right) /2 \\ 0, \quad otherwise \end{cases} \tag{6.1}$$

sachant que :

— Equation 6.2 : Degré externe normalisé

$$degree\,(u,v)_{norm} = \frac{degree\,(u,v)}{\min\left(|u|,|v|\right)} \tag{6.2}$$

— Equation 6.3 : Degré externe de $egonet(u)$ à $egonet(v)$

$$degree\,(u,v) = \left|V^{egonet(u)} \cap V^{egonet(v)}\right| + \left|uv \in E : u \in V^{egonet(u)}, v \in V^{egonet(v)}\right|, \quad u,v \in G \tag{6.3}$$

D'où $V^{egonet(u)}$ est l'ensemble des noeuds de $egonet(u)$ et $V^{egonet(v)}$ est l'ensemble des noeuds de $egonet(v)$.

6.2.3 Estimation du score d'anomalie total

Dans cette section, nous développons une nouvelle formule d'estimation d'un score d'anomalie total compris entre 0 et 1 pour chaque nœud du réseau multidimensionnel afin qu'il nous aide à prendre la bonne décision concernant la nature du comportement de l'utilisateur. Tout d'abord, nous calculons le score d'anomalie $AS(u)$ de chaque nœud dans chaque dimension d_i. Ensuite, nous calculons deux autres scores $DE(u)$ et $nbct(u)$ afin de spécifier l'influence du nœud u sur les nœuds appartenant à sa communauté. Enfin, un score total d'anomalie $AST(u)$ est estimé.

— Etape 1 : Nous commençons par calculer le score d'anomalie de chaque nœud dans chaque dimension existante dans notre réseau multidimensionnel. Le nœud peut avoir un score de : 0, 1 ou 0,5. Ce score est attribué en fonction de l'influence du nœud sur sa communauté (voir l'équation 6.4).

$$AS(u)_{d_i} = \begin{cases} 1,\ si\ \ (u \in Com)\, et\, (u\ influence\ fortement\ la\ construction \\ \qquad\qquad de\ la\ communauté) \\ 0.5,\ si\ \ (u \in Com)\, et\, (u\ \ n'influence\ pas\ la\ construction \\ \qquad\qquad de\ la\ communauté) \\ 0,\ si\ \ (u \notin Com)\, et\, (u \in d_i) \end{cases} \tag{6.4}$$

Une question qui se pose dans le premier cas : comment pouvons-nous décider qu'un nœud influence fortement la construction de la communauté ? La réponse nécessite le calcul de deux scores. Le premier score $DE(u)$

indique la distance entre $ego(u)$ et $ego(v)$, et le second score $nbct(u)$ représente le nombre total de liens directs à partir de $ego(u)$ à $ego(v)$. Les équations 6.5, 6.6 et 6.7 indiquent respectivement la manière de calculer les trois scores $DE(u)$, $nbct(u)$ et $nbc(u)$ dans chaque dimension.

$$DE(u)_{d_i} = nombre\ de\ liens\ sortants\ possibles\ du\ noeud(u)\ vers\ tous\ les\ noeuds\ du\ Com(u) - nombre\ de\ liens\ sortants\ du\ noeud(u)\ vers\ ses\ voisins\ directs \tag{6.5}$$

$$nbct(u)_{d_i} = \frac{\sum nbc(u)_{d_i}}{nombre\ de\ noeuds\ formants\ Com(u)} \tag{6.6}$$

Sachant que :

$$nbc(u)_{d_i} = nombre\ de\ liens\ directs\ entre\ l'egonet(u)\ et\ l'egonet(v) \tag{6.7}$$

La comparaison des deux scores $DE(u)$ et $nbct(u)$ peut nous permettre de déduire le degré d'influence du $noeud(u)$ sur les nœuds de sa communauté. D'où, si le score $DE(u)$ est supérieur ou égal au score $nbct(u)$, le $noeud(u)$ a une relation de degré moyen avec les nœuds appartenant à sa communauté, et si le score $DE(u)$ est inférieur au score $nbct(u)$, le $noeud(u)$ est fortement connecté aux nœuds qui constituent sa communauté (voir l'équation 6.8).

$$AS(u)_{d_i} = \begin{cases} si \quad (u \in Com) \quad alors \\ 1, \quad DE(u) < nbct(u) \\ 0.5, \quad DE(u) \geq nbct(u) \\ 0, \quad sinon \end{cases} \tag{6.8}$$

Etape 2 : Le calcul du score d'anomalie total de chaque nœud u est exprimé en fonction de la somme des scores d'anomalie du nœud u dans chaque dimension et du nombre de dimensions où le nœud u existe. L'équation 6.9 représente la formule permettant de calculer le score total d'anomalies pour chaque nœud.

$$AST(u) = \frac{\sum AS(u)_{d_i}}{nombre\ de\ dimensions\ où\ (u)\ existe} \tag{6.9}$$

6.2.4 Classification automatique des comportements anormaux

Dans cette section, nous expliquons la méthode que nous avons adopté pour classifier automatiquement les anomalies basées sur les scores $AST(u)$ calculés lors de la phase précédente.

Dans la littérature, il existe deux solutions fréquentes au problème de la classification. Une première solution consiste à classer les scores $AST(u)$ par rang croissant en sélectionnant les premiers ou les derniers k nœuds. Cependant, la principale difficulté de cette solution consiste à spécifier la valeur de k appropriée pour tous les ensembles de données, ce qui peut entraîner une erreur. Une deuxième solution consiste à définir un seuil de séparation entre les scores $AST(u)$. L'identification d'un seuil approprié pour tous les types de données est une tâche difficile.

Pour remédier à cela, nous utilisons le modèle de mélange de la distribution Beta, qui est un moyen efficace de résoudre ce type de problème [133][134][135]. Il suffit alors de déterminer les conditions d'application du modèle de probabilité de mélange pour la reconnaissance de la nature du comportement.

La distribution Beta admet l'adaptabilité et la souplesse nécessaires pour modéliser des situations complexes et variables, contrairement à d'autres distributions statistiques [136]. Par exemple, la distribution gaussienne ne permet que de modéliser des modes symétriques, ce qui exprime la possibilité d'obtenir une modélisation moins adéquate des données [137]. La distribution Beta se caractérise également par la modélisation sous différentes formes : la forme en U, la forme en L, la forme d'une droite [136], ce qui montre sa forte adaptabilité pour la modélisation précise de nos scores d'anomalie.

La phase de détection automatique des anomalies nécessite l'application de deux algorithmes : (1) l'algorithme d'estimation du nombre optimal de composantes (p) de la distribution Beta et (2) l'algorithme d'identification automatique de noeuds anormaux [27]. Le choix de l'application de ces deux algorithmes présente deux avantages :

— Le premier avantage consiste à identifier les anomalies par une méthode adéquate basée sur une loi de probabilité.

— Le deuxième avantage est l'utilisation de la même méthode de classification des scores d'anomalie que le travail [27]. Ce travail est le travail unique de détection d'anomalies basé sur des graphes multidimensionnels utilisant une telle classification. Compte tenu du but commun de nos travaux, une étude comparative peut avoir lieu.

Les deux algorithmes mentionnés précédemment peuvent être résumés en trois

étapes : (1) l'estimation des paramètres d'un composant, (2) l'application de l'algorithme Expectation Maximization (EM) pour la distribution Beta et (3) l'estimation du nombre optimal de composants.

6.2.4.1 Initiation au modèle de la distribution Beta

En théorie des probabilités et des statistiques, la loi Beta est une famille de lois de probabilités continues, définies sur l'intervalle [0,1], paramétrées par deux paramètres de forme, généralement notées α et β. Dans notre méthode, les scores $AST(u)$ estimés lors de la phase précédente sont compris entre 0 et 1. Ces scores sont répartis selon la loi Beta $AST \sim Be(\alpha;\beta)$, d'où sa densité de probabilité suit l'équation 6.10 :

$$B(AST) = \frac{1}{Be(\alpha;\beta)} AST^{\alpha-1}\,(1-AST)^{\beta-1} \tag{6.10}$$

sachant que la fonction Beta $Be(\alpha;\beta)$ est définie par l'équation 6.11 :

$$Be(\alpha;\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{6.11}$$

La fonction Γ est la fonction Gamma définie par l'équation 6.12 :

$$\Gamma : z \to \int_0^{+\infty} t^{z-1}\exp^{-t}\,dt \tag{6.12}$$

6.2.4.2 Estimation des paramètres d'un composant

Pour estimer les paramètres α et β d'une composante de la distribution Beta, il est essentiel de calculer la moyenne empirique de notre échantillon (N : le nombre de nœuds dans le graphe) et la variance d'un composant. Soit $\overline{x}$ la moyenne empirique de l'échantillon défini par l'équation 6.13 :

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} AST_i \tag{6.13}$$

et v la variance d'un composant défini par l'équation 6.14 :

$$v = \frac{1}{N}\sum_{i=1}^{N} \left(AST_i - \overline{x}\right)^2 \tag{6.14}$$

Les estimations de α et β sont les suivantes (voir les équations 6.15 et 6.16) :

$$\alpha = \overline{x}\left(\frac{\overline{x}\left(1-\overline{x}\right)}{v} - 1\right) \tag{6.15}$$

$$\beta = \left(1-\overline{x}\right)\left(\frac{\overline{x}\left(1-\overline{x}\right)}{v} - 1\right) \tag{6.16}$$

6.2.4.3 Application de l'algorithme EM pour la distribution Beta

Comme son nom l'indique, l'approche Maximum Likelihood Estimation (MLE) consiste à maximiser la vraisemblance, c'est-à-dire maximiser $L(\Theta, AST) = \prod_{i=1}^{N}\sum_{k=1}^{k}\beta_k B_k\left(AST_i;\alpha_k\right)$ ou maximiser de manière équivalente le log de la vraisemblance $l(\Theta, AST) = \sum_{i=1}^{N}\log\left(\sum_{k=1}^{k}\beta_k B_k\left(AST_i;\alpha_k\right)\right)$ afin d'estimer les paramètres inconnus, avec $\Theta = (\alpha_1, \beta_1, ..., \alpha_k, \beta_k)$ les paramètres inconnus du modèle paramétrique.

Cependant, ce problème de maximisation ne peut pas être résolu de manière analytique en raison de données cachées. Nous devons trouver une solution en utilisant des algorithmes itératifs. Parmi ces algorithmes, nous citons l'algorithme EM [138]. Cet algorithme a pour but de fournir un estimateur lorsqu'il est impossible de calculer la solution en raison de la présence de données cachées ou manquantes ou plutôt lorsque la connaissance de ces données permettrait l'estimation des paramètres.

L'algorithme EM opère deux étapes distinctes, à savoir :

— la phase "Expectation", souvent appelée "étape E", procéde à l'estimation des données inconnues, en connaissant les données observées et la valeur des paramètres déterminée à l'itération précédente ;

— la phase "Maximisation", ou "M", procède à la maximisation de la vraisemblance, rendue désormais possible en utilisant l'estimation des données inconnues effectuée à l'étape précédente, et met à jour la valeur du ou des paramètre(s) pour la prochaine itération.

Cet algorithme garantit que la vraisemblance augmente à chaque itération, ce qui conduit à des estimateurs de plus en plus précis. Plus formellement et plus généralement dans le cas discret :

— Si nous disposons de l'observation $AST = (AST_1, ..., AST_N)$ de vraisemblance dénotée par $P(AST \backslash \Theta)$ dont la maximisation est impossible.

— Nous considérons les données cachées $Z = (Z_1, ..., Z_N)$ dont la connaissance permettrait de maximiser "la probabilité de données complètes" notée par $P(AST, Z \backslash \Theta)$.

— Comme ces données Z sont inconnues, nous estimons la probabilité que les données complètes prennent en compte toutes les informations connues : pour cela, nous avons choisi comme estimateur $E_{Z \backslash AST, \Theta_m}[\log P(AST, Z \backslash \Theta)]$; (c'est "l'étape E" de l'algorithme)

— Nous maximisons enfin cette probabilité estimée pour déterminer la nouvelle valeur du paramètre ("étape M" de l'algorithme).

Ainsi, la transition de l'itération m à l'itération $m + 1$ de l'algorithme consiste à déterminer : $\Theta_{m+1} = argmax \;\; E_{Z \backslash AST, \Theta_m}[\log P(AST, Z \backslash \Theta)]$.

6.2.4.4 Estimation du nombre de composantes (p) de la distribution Beta

Pour trouver le bon modèle pour nos données, nous devons estimer le nombre de composants p et les paramètres α et β de chaque composant. Le nombre de composants p varie entre 1 et $p - max$. Pour chaque composant calculé, les métriques de performance sont déterminées pour identifier le nombre optimal de composants. Pour ce faire, le Bayesian Information Criterion (BIC) a été utilisé [139]. Le critère BIC s'écrit comme suit (Equation 6.17) :

$$BIC(p) = -2\log(L_p) + k_p \log(N) \tag{6.17}$$

avec L : la probabilité du modèle estimé, N : le nombre d'observations dans l'échantillon et k : le nombre total de paramètres estimés du modèle.

Le modèle qui sera sélectionné est celui qui minimise le critère BIC (voir l'équation 6.18) :

$$M_{BIC_p} = arg_{min}\ BIC_p(M) \tag{6.18}$$

6.3 Implémentation et évaluation des résultats

Cette section présente une évaluation empirique de la performance de notre méthode sur un réseau tridimentionnel. Dans ce contexte, nous élaborons un ensemble d'expérimentations pour évaluer de façon objective l'applicabilité de notre méthode. Nous commençons d'abord par décrire le cadre expérimental. Par la suite, nous présentons une description détaillée des données réelles testées. Et enfin, nous interprétons la pertinence des résultats obtenus.

Les différentes dimensions du réseau tridimentionnel sont les suivantes :(1) Facebook, (2) Twitter et (3) Instagram. Nous n'avons aucune connaissance à priori sur le partitionnement des nœuds, c.-à-d., une connaissance au préalable de l'appartenance des nœuds à une catégorie spécifique, à savoir, anomalie ou nœud normal, est manquante. De ce fait, nous ne pouvons pas utiliser des métriques supervisées qui se basent sur l'existence d'un partitionnement de référence. Dans ce contexte, nous avons adopté une approche objective qui consiste à l'interprétation des interactions des nœuds par une investigation manuelle et une visualisation graphique de la matrice d'adjacence du réseau.

Pour chaque noeud du réseau, nous avons estimé un score d'anomalie total. Nous avons modélisé, ensuite, la distribution de ces scores selon notre modèle probabiliste qui exploite la distribution Beta. Cela est représenté à partir de la courbe de densité des scores d'anomalies du réseau étudié (voir la figure 6.1). La courbe de densité nous a permis de constater la grande flexibilité et l'adaptabilité du modèle Beta à modéliser les distributions. Dans la figure 6.1, le premier composant (proche de zéro) représente les valeurs des scores d'anomalies les plus faibles. Par conséquent, les nœuds associés aux scores qui sont groupés dans ce composant sont identifiés comme des anomalies.

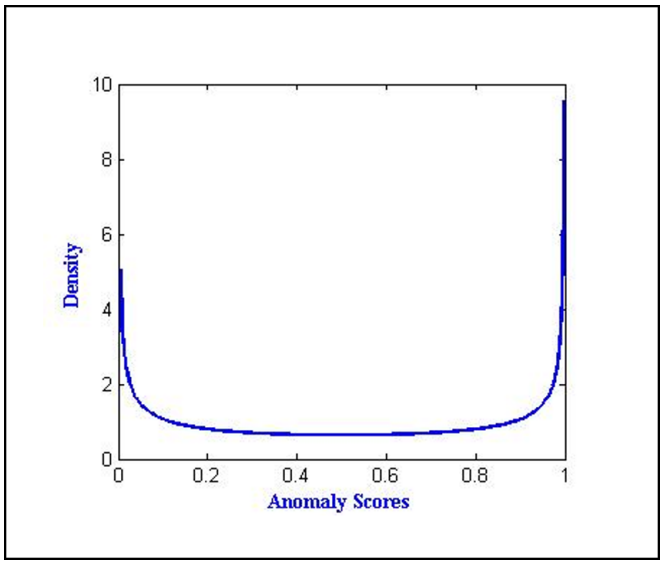

Figure 6.1: Densité de la probabilité des scores estimés.

En ce qui concerne le réseau testé, environ 10K de nœuds d'un ensemble de 397K ont été sélectionnés comme des nœuds ayant des connexions atypiques. La figure 6.2 présente la matrice d'adjacence des trois dimensions du réseau de sorte que les nœuds soient triés suivant un ordre ascendant de leurs scores d'anomalie. Ayant les scores $AST(u)$ les plus faibles, les anomalies sont placées en haut de la matrice. Par conséquent, ces anomalies sont connectées d'une façon éparse sur le réseau, par contre les nœuds normaux sont étroitement connectés et cela se manifeste par les régions denses.

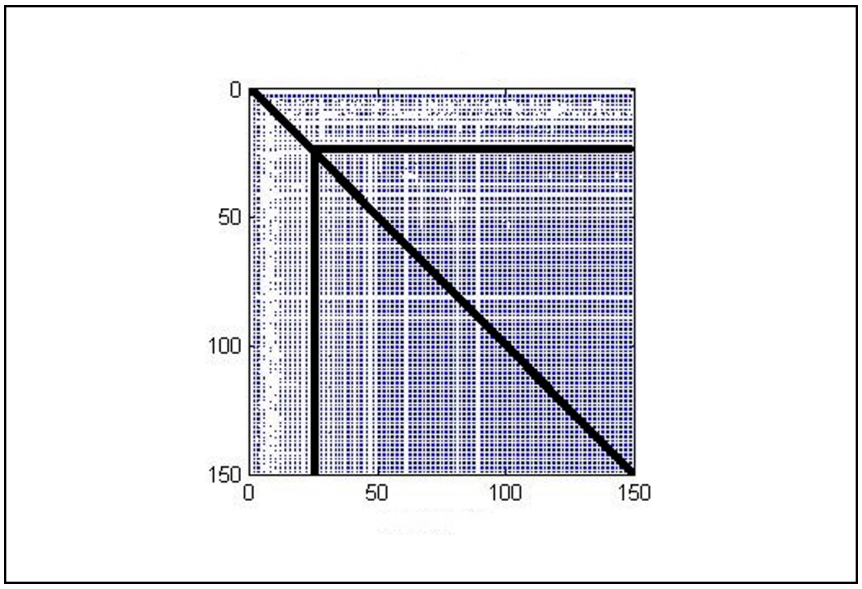

Figure 6.2: Matrice d'adjacence du réseau étudié.

6.4 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle méthode de détection d'anomalies sur la base de l'analyse des relations entre les utilisateurs dans une structure multidimensionnelle. L'approche proposée procède en trois phases. Dans la première phase, nous avons construit des communautés utilisateurs afin de pouvoir identifier et justifier la force de connexion entre les utilisateurs dans différentes dimensions. Dans la deuxième phase, nous avons estimé un score d'anomalies pour chaque utilisateur pour connaitre sa position par rapport aux autres utilisateurs. Dans la troisième phase, nous avons utilisé une méthode probabiliste qui se base sur la distribution Beta afin de classifier les scores d'anomalies. Nous avons, par la suite, évalué l'approche proposée sur un réseau tridimentionnel réel. Les résultats obtenus ont montré la validité de notre méthode. Ce travail a été publié dans les actes de la conférence AINA 2020 [1]. Dans le chapitre suivant, nous étendons cette

1. https ://doi.org/10.1007/978-3-030-44041-1_55

méthode afin d'améliorer ses performances en combinant l'aspect comportemental et structurel qui explore la structure topologique multidimentionnelle pour analyser les relations et les activités des utilisateurs à la fois en manipulant plus de types de données que les contributions précédentes.

Chapitre 7

Framework hybride de détection des comportements anormaux sur un réseau multidimensionnel utilisant des données multimodales

Sommaire

7.1 Introduction

Les agences de renseignement gouvernementales (par exemple le FBI) utilisent une variété de tactiques et de stratégies pour lutter contre le terrorisme. Ils collectent des informations afin de garantir la sécurité de diverses manières : interception de communications cryptées, analyse d'informations publiques, actions de surveillance discrète, contacts et paroles de personnes. De nos jours, cette mission est devenue un peu plus compliquée étant donné l'imprévisibilité du comportement terroriste. Les terroristes sont de plus en plus prudents dans leurs déplacements, notamment dans l'utilisation des réseaux sociaux. Pour faire face à ce problème, plusieurs travaux récents ont montré leur intérêt à l'identification d'utilisateurs terroristes dans des environnements collaboratifs en ligne. Dans ce chapitre, nous étendons les deux contributions élaborées dans le chapitre 5 et chapitre 6 afin de

proposer un nouveau framework hybride de détection des comportements terroristes capable d'analyser une structure sociale multidimensionnelle, comportementale et structurelle. Ce dernier prend en entrée un type de données nouveau par rapport aux deux contributions précédentes, à savoir, les informations générales de l'utilisateur. L'objectif de cette extension est de prouver que l'adaptation de plusieurs dimensions de réseaux sociaux ainsi que de divers types de données permet d'augmenter de loin la précision de l'analyse et de la détection.

7.2 Méthodologie générale

Dans cette section, le principe général du framework est expliqué en détail. Dans un premier temps, un réseau multidimensionnel est construit à partir d'un ensemble de données réelles provenant de différents réseaux sociaux (Facebook, Instagram et Twitter). Ensuite, la méthode de détection des comportements anormaux présentée dans le chapitre 6 est appliquée. Enfin, un nouveau modèle est alimenté par plusieurs types de données (textes, images et informations générales) provenant de diverses sources (en ligne et hors ligne) sur le terrorisme. Ce modèle est conçu pour permettre l'identification des comportements terroristes. La figure 7.1 montre les différentes étapes de notre méthodologie.

7.2.1 Collecte de données

Nous nous concentrons dans cette section sur l'extraction de trois types de données : (1) les données textuelles, qui comprennent les messages, les commentaires, les légendes d'images, le texte dans une image, etc., (2) les données de type image, comme par exemple les photos publiées, les photos de profil, etc. et (3) les données de type numérique, comme l'âge, le nombre d'amis, le nombre moyen de messages par jour, etc. Plusieurs autres informations existent dans les médias

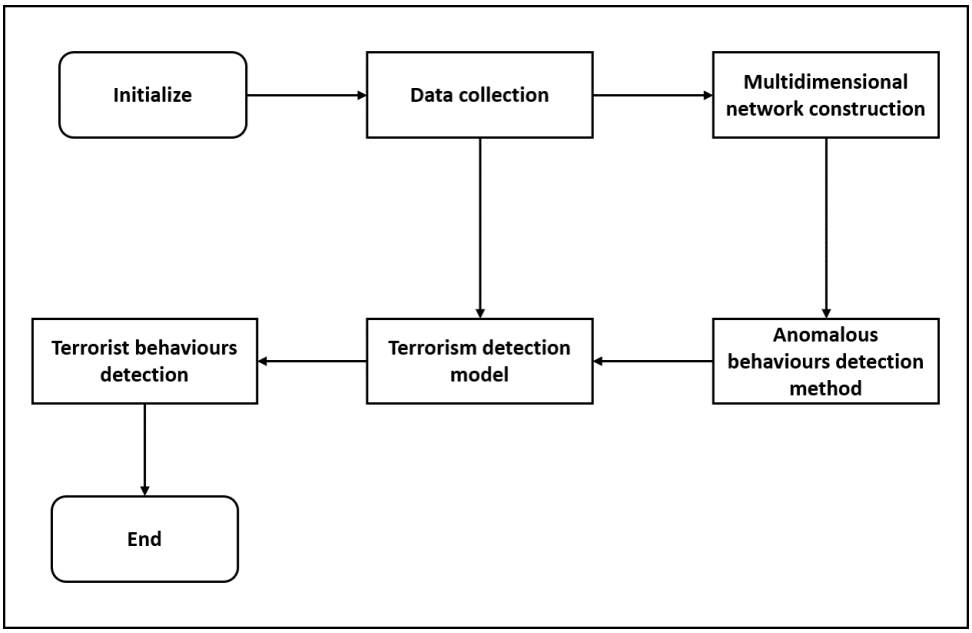

Figure 7.1: Principe général du framework

sociaux, comme le nom d'utilisateur, le sexe, la ville natale, la ville de résidence, les membres de la familles, etc. Par conséquent, au lieu d'avoir une catégorie de données de contenu numérique, nous avons opté pour l'utilisation d'une autre catégorie appelée "Données d'informations générales" qui comprend les données de contenu numérique et les données d'information de l'utilisateur. Dans cette section, nous présentons les sources de données hors ligne et en ligne qui sont utilisées pour récupérer nos types de données cibles pour l'entraînement du modèle et la détection ultérieure d'événements terroristes. Afin d'entraîner le modèle et de distinguer précisément les profils de terroristes et de réduire les faux positifs (des profils qui ressemblent à des profils terroristes sans l'être), nous avons décidé de considérer (comme c'était le cas dans la première contribution) les contenus terroristes comme des étiquettes positives et les contenus militaires et d'actualité comme des étiquettes négatives. Comme ces types de contenus sont liés, l'entraînement de ces derniers les uns par rapport aux autres rendra le modèle plus précis.

7.2.1.1 Sources de données hors ligne

Les données hors ligne sont les données utilisées pour l'entraînement du modèle, que nous avons recueillies à partir de jeux de données publics sur le terrorisme. Pour chaque type d'entrée, nous avons utilisé un ensemble de données différent. Pour les données de contenu textuel, nous nous sommes inspirés de [28] pour utiliser l'API de Twitter afin de rassembler les tweets qui contiennent des hashtags liés au terrorisme et les tweets des comptes terroristes qui ont été signalés au compte de sécurité de Twitter (twittersafety) en s'assurant qu'il ne s'agit pas de comptes anti-terroristes. De cette façon, nous avons crée notre ensemble de données de contenu textuel hors ligne et nous avons considéré ces tweets comme des étiquettes positives. Alors que les tweets d'actualité terroriste et les titres d'actualité recueillis dans d'autres ensembles de données publiques tels que la base de données mondiale sur le terrorisme (GTD) [113] ont été considérés comme des étiquettes négatives. Nous avons également utilisé l'API google translate car certains comptes peuvent publier des tweets dans différentes langues. Pour les données de type image, nous n'avons pas trouvé d'ensemble de données publiques d'images liées au terrorisme dans le cadre de notre recherche. Nous avons décidé d'utiliser une méthode de grattage manuel du Web avec Google Image comme source de données. Nous avons rassemblé manuellement les images d'individus terroristes et les images d'incitation au terrorisme, qui sont nos étiquettes positives, et nous les avons opposées aux images d'actualités militaires et terroristes, qui sont nos étiquettes négatives. Pour les données d'information générales, Study of Terrorism And Responses to Terrorism (START) a publié une base de données appelée "Profiles of Individual Radicalization In the United States" (PIRUS) [1], qui contient

1. https ://www.start.umd.edu/data-tools/profiles-individual-radicalization-united-states-pirus

environ 145 caractéristiques sur de nombreux profils radicaux dont nous avons extrait les caractéristiques pertinentes pour notre travail, à savoir l'âge, le sexe, les relations, etc.

7.2.1.2 Sources de données en ligne

Les données en ligne sont les données des réseaux sociaux qui font partie de la détection et du réentraînement du modèle futur. Pour les médias sociaux, nous avons décidé d'étudier trois sites Web populaires qui ont des contenus de données similaires et qui peuvent être reliés entre eux : Facebook, Instagram et Twitter. Facebook fournit un Graph API [2], un service API basé sur HTTP permettant d'accéder aux objets du graphe social de Facebook [3]. Les données sont riches en sémantique puisque le Graph API utilise le format Resource Description Framework (RDF) comme type de retour [140]. Instagram, qui fait partie de Facebook, fournit également une API graphique pour les comptes professionnels [4]. Pour les comptes d'utilisateurs normaux, il fournit une API REST qui renvoie un objet JSON pour l'interrogation des données publiques [5]. Twitter met à disposition une API REST [6] avec un format de retour JSON qui fournit plusieurs requêtes de données publiques ainsi que des données privées avec les autorisations appropriées [141].

2. https ://developers.facebook.com/docs/graph-api
3. https ://www.cbsnews.com/news/facebook-one-social-graph-to-rule-them-all/
4. https ://developers.facebook.com/docs/instagram-api
5. https ://www.instagram.com/developer/
6. https ://developer.twitter.com/en/docs.html

7.2.1.3 Données de test

Des données de test réelles sont extraites des trois réseaux sociaux mentionnés dans la sous-section précédente afin de montrer les performances de notre framework. Un premier ensemble comprend les données Facebook [7] de 180 personnalités publiques célèbres actives dans divers domaines (politique, sport, culture, etc.). Cet ensemble contient les posts de ces personnes qui datent du 24 février 2016 et qui sont obtenus par un scraper. Un second jeu de données est issu de Twitter et qui enregistre les 200 derniers posts des mêmes personnages considérés avec Facebook. Ces données ont été collectées par l'outil de scraping Octoparse et enrichies par un scraper, que nous avons développé. De même, un troisième ensemble de données agrégées provenant d'Instagram a été obtenu. Ces trois ensembles sont liés par les identifiants des personnes afin de garantir la synchronisation entre les données collectées. Un graphe multidimensionnel a ensuite été modélisé en ajoutant les différentes relations entre les nœuds qui représentent les personnes.

7.2.2 Méthode de détection d'anomalies

Dans cette sous-section, nous expliquons brièvement les étapes suivies pour réaliser l'identification des nœuds atypiques dans une structure multidimensionnelle. Notre méthode, présentée dans le chapitre précédent (Chapitre 6), fonctionne principalement en trois phases :

— Phase 1 : la détection des différentes communautés d'utilisateurs suivant l'utilisation d'une formule basée sur le calcul du degré externe normalisé de chaque pair de nœuds.

— Phase 2 : l'estimation du score d'anomalie pour chaque nœud du graphe en considérant le degré d'influence du nœud dans sa communauté.

7. https ://github.com/minimaxir/interactive-facebook-reactions

— Phase 3 : la classification des différents scores obtenus par le modèle de mélange Beta. Ce choix d'un classifieur s'explique par l'efficience de ce modèle probabiliste pour la détection de nœuds anormaux. L'application d'une telle méthode a permis d'obtenir un ensemble de nœuds présentant un caractère anormal. Un modèle de détection d'un comportement spécifique entraîné sur des données terroristes analysera ces nœuds dans ce qui suit.

7.2.3 Modèle hybride de détection du terrorisme

Dans cette sous-section, nous présentons un nouveau modèle de détection des comportements terroristes. Il considère trois sous-modèles différents : un sous-modèle pour chaque type de données d'entrée (modèle d'analyse de texte, modèle d'analyse d'image et modèle d'analyse d'informations générales). Le workflow de notre modèle est illustré par la figure 7.2. Chaque fois qu'un utilisateur est impliqué dans une activité, les données de l'utilisateur passent à nouveau par notre modèle. Si l'utilisateur est détecté comme terroriste, nous reformons le modèle avec ces nouvelles données pour le maintenir à jour avec les nouveaux comportements invisibles. Si à la fin de l'entraînement le modèle perd de sa précision, nous revenons au dernier modèle existant.

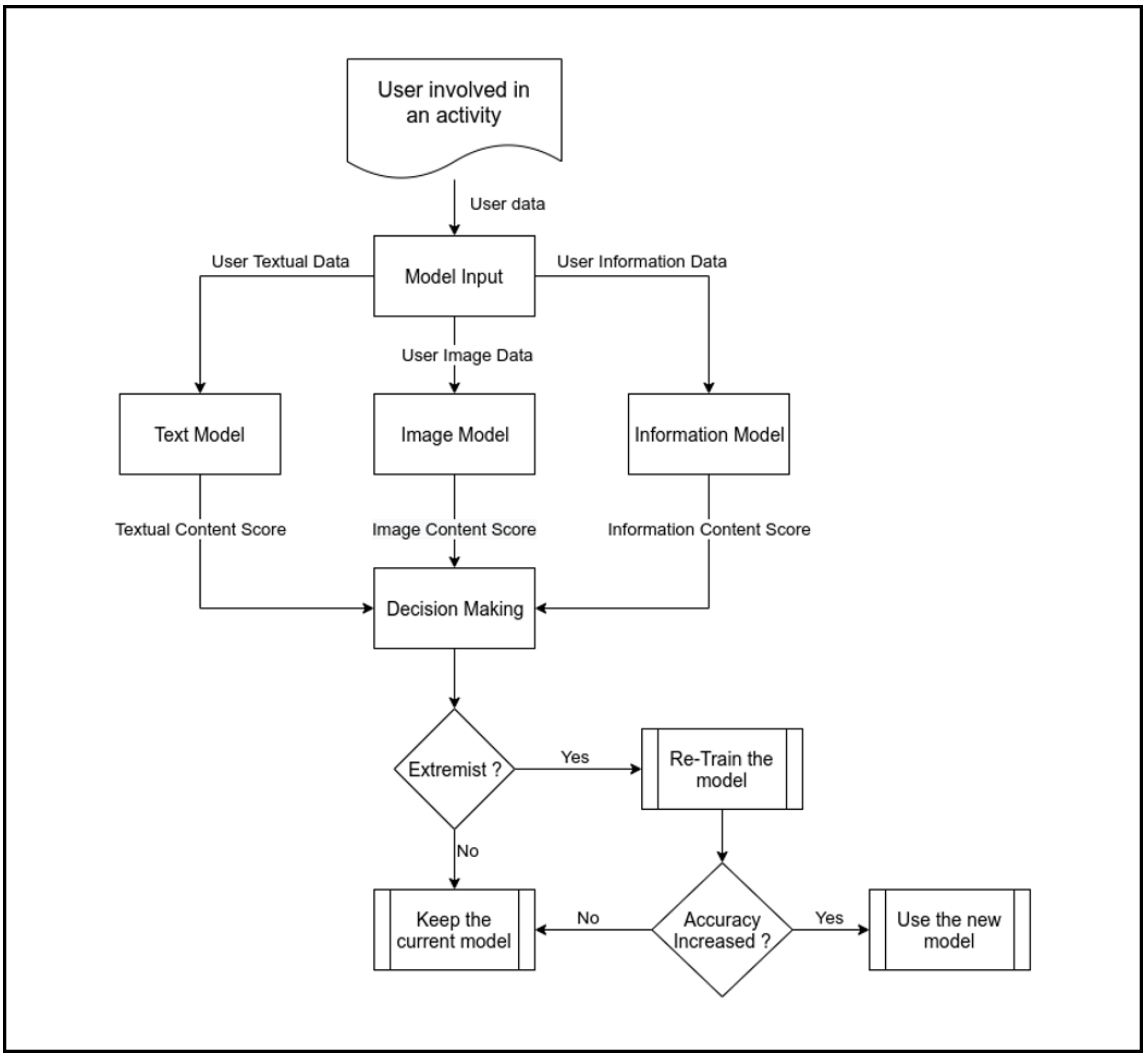

Figure 7.2: Workflow de notre modèle.

7.2.3.1 Modèle de classification de texte

Comme l'illustre la figure 7.3, dès que les données textuelles sont reçues, elles passent par le processus NLP. Ensuite, elles sont converties en valeurs numériques. Pour cela, nous avons comparé deux techniques d'incorporation de mots, à savoir : Word2Vec [80] et TF-IDF [81]. Nous avons choisi TF-IDF parce que nous avons affaire à un problème de classification et que nous sommes plus intéressés par la

distinction des catégories que par la similarité des mots (le même choix retenu avec la première contribution). Après la conversion, les données sont prêtes à être traitées par n'importe quel modèle d'apprentissage automatique. Comme stratégie, nous avons décidé que dans la phase de mise en œuvre, nous allons tester différents modèles tels que SVM, LR et NN, puis comparer leurs résultats pour évaluer lequel est le plus performant.

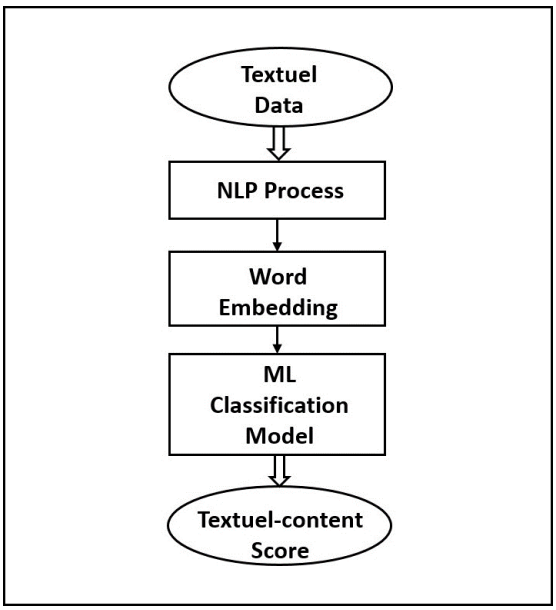

Figure 7.3: Modèle de classification de texte.

7.2.3.2 Modèle de classification d'image

Comme nous l'avons évoqué dans le chapitre 5, plusieurs travaux existants ont montré l'efficacité des CNN pour la classification d'images [83]. Cependant, la conception d'un CNN nécessite un grand nombre de réglages de paramètres et l'ajout/suppression de blocs de convolution pour trouver la meilleure architecture tout en ré-entraînant notre modèle à chaque fois. Cette tâche prend énormément de temps. Pour surmonter ce problème, la technique d'apprentissage TL permet

de ré-entraîner le modèle très rapidement [8]. Une autre limitation connue que nous rencontrons généralement dans la classification d'images est le manque de diversité des données et des échantillons. Pour surmonter ces problèmes, la technique DA, retenue aussi au sein du chapitre 5, génère des données plus variées pour améliorer la précision des résultats [87]. Le processus global de ce sous-modèle est décrit dans la figure 7.3.

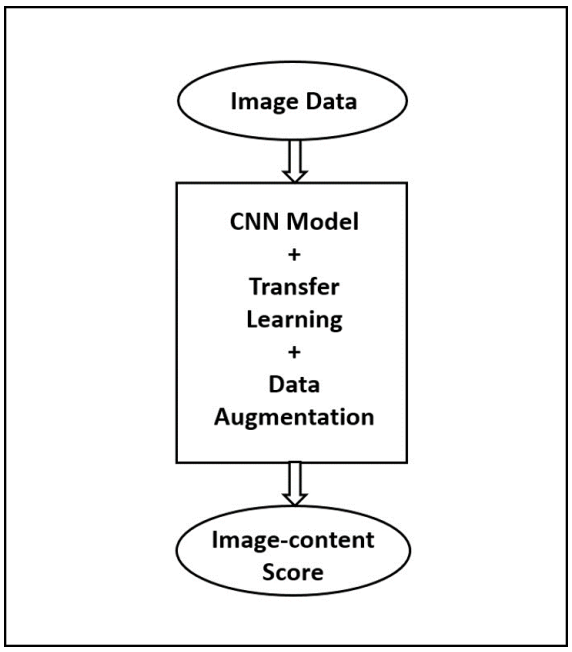

Figure 7.4: Modèle de classification d'image.

7.2.3.3 Modèle de classification d'informations générales

Comme l'illustre la figure 7.4, les caractéristiques de ce modèle ne nécessitent pas de traitement préalable pour que la machine puisse les comprendre. Les caractéristiques non numériques, telles que le sexe et la relation, doivent être encodées

8. google_images_download : Python Script to download hundreds of images from 'Google Images'. It is a ready-to-run code ! URL https ://github.com/hardikvasa/google-images-download.

en valeurs numériques. Nous utilisons le codage 0/1 pour les caractéristiques binaires et les codages "One-hot" ou "Sparse Categorical Cross Entropy" pour les valeurs non binaires. Pour le nom d'utilisateur, nous extrayons des informations utiles telles que la longueur, le nombre de caractères uniques et d'autres informations importantes. D'autres caractéristiques numériques telles que l'âge, le nombre d'amis et le nombre de followers peuvent être récupérées directement dans le modèle. Dans la phase d'implémentation, nous avons essayé différents modèles de classification et nous avons comparé leurs résultats pour sélectionner le meilleur.

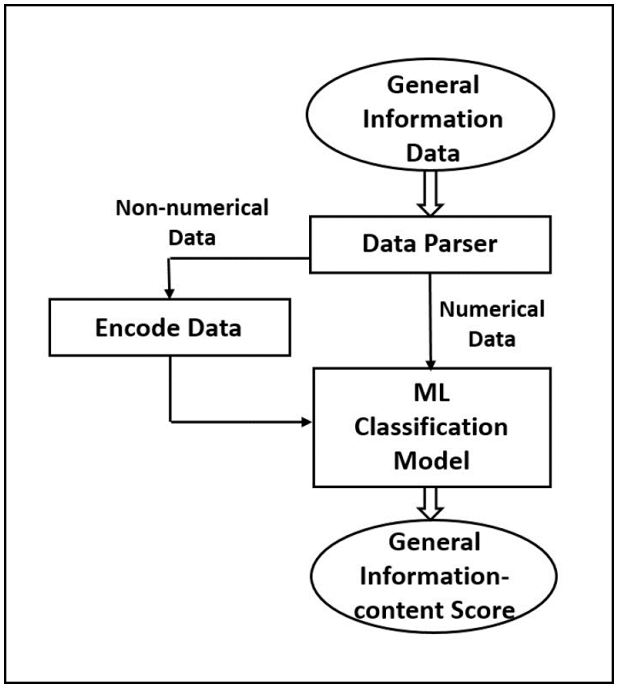

Figure 7.5: Modèle de classification d'information générale.

7.2.3.4 Composante décisionnelle

La composante décisionnelle suit la même démarche que celle du composant de prise de décison abordé dans le chapite 5. Elle est chargée de fournir un score global pour chaque utilisateur lors de l'interconnexion des médias sociaux. Lors de nos

expériences, et sur la base des fonctionnalités disponibles, nous avons remarqué que les contenus textuels et les images ont plus d'impact sur le comportement de l'utilisateur que les informations générales qui peuvent être trompeuses.

7.3 Résultats expérimentaux

7.3.1 Collecte de données

Il est à signaler que l'une des difficultées majeures dans ce travail est la ddisponibilité des données. Pour surmonter cette difficulté, nous avons eu recours à plusieurs sources et moyens d'extraction de données. Nous n'avons collecté que des données publiques sur le terrorisme en raison de la confidentialité et de l'inexistence de ce type de données, compte tenu des règles de fermeture de compte appliquées par les réseaux sociaux. Une quantité considérable de données a été collectée et analysée à des fins de lutte contre le terrorisme.

7.3.1.1 Collecte de données hors ligne

Données textuelles. Elles sont collectées à partir de comptes de tweeter interdits (Labels positifs) et de titres de nouvelles du GTD (Labels négatifs) comme expliqué dans le tableau 7.1.

Données de type image. Comme nous l'avons prémentionné dans notre recherche, la source des données d'image est Google-Image, et nous allons recueillir manuellement les images à partir de celui-ci. Heureusement, il existe une librairie Python appelée « google images download », qui nous permet d'automatiser cette tâche en choisissant les mots clés que nous recherchons et le nombre d'images nécessaires. Nous avons développé un script qui a téléchargé environ 500 images de personnes terroristes et d'actes d'incitation au terrorisme, en plus de 500 autres

images d'actualités militaires et de terrorisme. Malheureusement, ces images ne correspondaient pas à 100% à ce que nous recherchions ; nous avons donc dû vérifier manuellement les images recueillies et supprimer les images sans rapport. Après avoir nettoyé les données et gardé uniquement les images liées, nous avions environ 200 images de terroristes (étiquettes positives) et 300 images de militaires et de nouvelles (étiquettes négatives), comme illustré dans le tableau 7.1.

Données d'informations générales Pour les données d'information générale, nous avons utilisé l'ensemble de données publiques PIRUS dont nous avons extrait l'âge, le sexe et le statut relationnel de 135 terroristes qui constituent nos étiquettes positives. Quant aux étiquettes négatives, nous avons utilisé les données en ligne pour construire notre ensemble de données.

Table 7.1: Ensembles de données à contenu textuel et à contenu d'image

Label	Echantillons de texte	Echantillons d'image
Labels positifs	122619	219
Labels négatifs	181691	314
Total	304310	533

7.3.1.2 Collecte de données en ligne

Données Facebook. Facebook fournit une API basée sur HTTP appelée Graph API. Un SDK public appelé "facebook-sdk" nous a aidé à écrire un script automatisé de collecte de données Facebook en utilisant Python.

Données Instagram. Pour Instagram, la tâche est plus facile car il fournit une API REST normale avec une sortie JSON où l'accès à chaque endpoint est direct via n'importe quel module de requête HTTP. En Python, nous avons utilisé le

module "requests" avec le endpoint Instagram [9] où nous avons pu accéder aux informations de l'utilisateur [10] et aux ses posts [11].

Données Twitter. Comme pour Instagram, Twitter fournit également une API REST, et met à disposition un SDK Python qui facilite l'utilisation de l'API. Pour l'utiliser, nous avons transmis quatre clés d'accès : la clé du consommateur, le secret du consommateur, la clé du jeton d'accès et le secret du jeton d'accès. Chaque clé possède des autorisations pertinentes qui permettent d'accéder soit aux données privées de l'utilisateur, soit aux données publiques de Twitter.

7.3.2 Modèles d'apprentissage

7.3.2.1 Modèle de classification de texte

Nous avons essayé trois modèles de classification, à savoir : LR, SVM et NN. Pour mettre en œuvre la LR et le SVM, nous avons utilisé la bibliothèque Scikit-Learn. Nous avons entraîné ces deux modèles avec les valeurs des paramètres suggérés par défaut. Pour le NN, nous avons utilisé Keras, un framework qui fonctionne au-dessus de TensorFlow. L'architecture de notre modèle est composée de trois couches, avec respectivement 16 neurones, 8 neurones et 1 neurone. Les deux premières couches ont une activation "relu" car cette activation a montré ses performances, et la dernière couche a une activation "sigmoïde" car c'est notre couche de sortie et nous traitons de plus un problème de classification binaire. Le modèle est compilé avec une "binary cross-entropy" comme fonction de perte et "adam" comme optimiseur. Pour les paramètres d'entraînement, nous avons utilisé 20 époques avec une taille de lot de 128 et 20% des données de validation extraites des données d'entraînement. Le tableau 7.2 illustre les scores métriques pour chaque modèle

9. https ://api.instagram.com/v1/
10. /users/self/ ?access_token={}
11. /users/self/media/recent/ ?access_token={}

ainsi que le temps d'exécution de l'entrainement. Ces modèles sont formés et testés avec les mêmes données sur la même machine. A la lumière de ces résultats d'évaluation, le modèle que nous avons sélectionné est le NN car il a le meilleur F1-score avec une bonne moyenne de temps de formation.

Table 7.2: Scores métriques du modèle textuel

Modèle	Précision	F1-Score	Temps d'entrainement
LR	0.9726	0.9674	39.9s
SVM	0.9626	0.9548	6h 48min 33s
NN	**0.9774**	**0.9719**	**1min 11s**

7.3.2.2 Modèle de classification d'image

Nous avons défini un CNN comme modèle de classification d'images ainsi que des techniques d'optimisation, à savoir TL et DA. Par conséquent, nous avons implémenté les fonctions DA, puis défini les connaissances acquises à utiliser dans le modèle initial. Dans le tableau 7.3, nous présentons les différents scores des combinaisons utilisant nos deux couches CNN, avec et sans le modèle pré-entraîné et avec et sans les données générées par DA. Bien que les scores aient été mesurés par les mêmes données de test, les données d'entrainement diffèrent lors de l'utilisation de DA. L'utilisation conjointe de DA et de TL a permis d'obtenir de meilleurs scores et un temps de formation moins long ; nous les avons donc utilisés dans notre modèle global.

7.3.2.3 Modèle de classification des informations générales

Pour le modèle d'informations générales, nous avons suivi la même stratégie que celle utilisée pour la classification des textes, à savoir tester trois modèles de classification, à savoir LR, SVM et NN, afin de choisir le meilleur. Pour LR et SVM,

Table 7.3: Scores métriques du modèle d'image

Modèle	Précision	F1-Score	Temps d'entrainement
CNN	0.7631	0.7219	3min 50s
CNN + DA	0.7781	0.7463	4min 12s
CNN + TL	0.8291	0.8103	8min 48s
CNN + DA + TL	**0.8571**	**0.8454**	**9min 23s**

nous avons utilisé les valeurs par défaut des paramètres de Scikit-Learn. Quant au NN, nous avons utilisé une architecture de quatre couches avec 16 neurones, 8 neurones, 4 neurones et 1 neurone respectivement. Une activation "relu" est utilisée pour les trois premières couches et une activation "sigmoïde" pour la dernière couche. Le modèle est compilé avec une "binary cross-entropy" comme fonction de perte et "adam" comme optimiseur. Pour les paramètres d'entraînement, nous avons utilisé 200 époques avec une taille de lot de 32 et 20% des données de validation extraites des données d'entraînement. Le tableau 7.4 illustre les scores métriques des modèles formés avec les mêmes données sur la même machine. Pour le modèle global, nous avons choisi le SVM, car il dépasse de loin les performances des autres modèles.

Table 7.4: Scores métriques du modèle des informations générales

Modèle	Précision	F1-Score	Temps d'entrainement
LR	0.7650	0.7873	5s
SVM	**0.8300**	**0.8495**	**7s**
NN	0.8173	0.8325	48.6s

7.3.2.4 Composante décisionnelle

La composante de prise de décision effectue les trois tâches suivantes :

— Tâche 1 : L'interprétation d'un score de sortie du modèle global en fonction des poids des sous-modèles. Pour calculer un poids pour chaque sous-modèle, nous nous sommes intéressés à l'extraction d'informations pertinentes afin de représenter les données par plusieurs caractéristiques. Ces caractéristiques sont présentées par les types de données utilisés lors de l'entraînement du modèle global, à savoir : les textes, les images et les informations générales. L'essentiel à ce stade est de calculer l'importance de l'impact de chacune de ces caractéristiques sur le modèle formé. Nous avons donc utilisé le modèle XGBoost [12] pour lequel la bibliothèque Python XGBoost propose plusieurs calculs. Le principe du modèle adopté est basé sur l'occurrence de la caractéristique dans le modèle formé. La figure 7.6 présente les occurrences de chaque caractéristique où il est très clair visible que les données textuelles ont un grand impact dans l'entrainement du modèle. La figure 7.7 montre les valeurs des poids calculés par le modèle XGBoost.

— Tâche 2 : Le calcul d'un score total de terrorisme dans toutes les dimensions du réseau.

— Tâche 3 : Le classement des scores globaux par la mise en place d'un seuil automatique. Comme les ensembles de données utilisés n'étaient pas étiquetés, nous avons utilisé l'inspection visuelle pour étiqueter de manière aléatoire 50 nœuds du graphe dans toutes les dimensions afin d'identifier un score de seuil aberrant qui minimise les faux négatifs et les faux positifs. Sklearn ne nous permet pas de définir directement le seuil de décision, mais il nous fournit tous les scores de décision calculés pendant la prédiction. Pour cette raison, nous avons implémenté une fonction automatique qui se charge de sélectionner le meilleur score de décision et de le définir comme valeur de seuil. Par la suite, pour plusieurs valeurs de seuil, notre fonction

12. https ://github.com/aureliale/features_importance

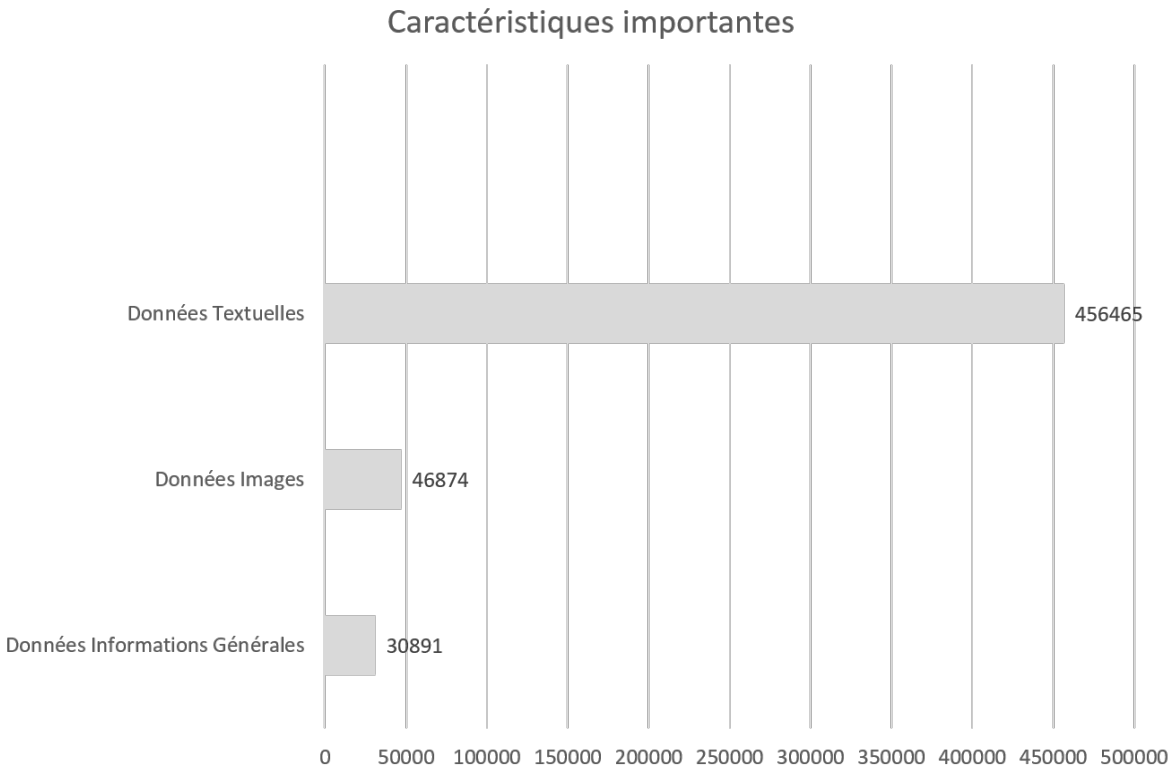

Figure 7.6: Caractéristiques importantes.

essaie de choisir la valeur de seuil qui augmenterait la précision et ne diminuerait pas trop le rappel. Par conséquent, les valeurs de score de décision qui sont inférieures à ce seuil de décision sont considérées comme une classe négative (0), et toutes les valeurs de score de décision qui sont supérieures à cette valeur de seuil sont classées comme une classe positive (1). Nous avons fini par choisir une valeur de 0,63 comme valeur seuil. Ainsi, notre fonction s'adapte automatiquement très bien avec les enrichissements de notre modèle, elle permet de calculer un nouveau score de seuil, qui convient le mieux aux nouvelles données.

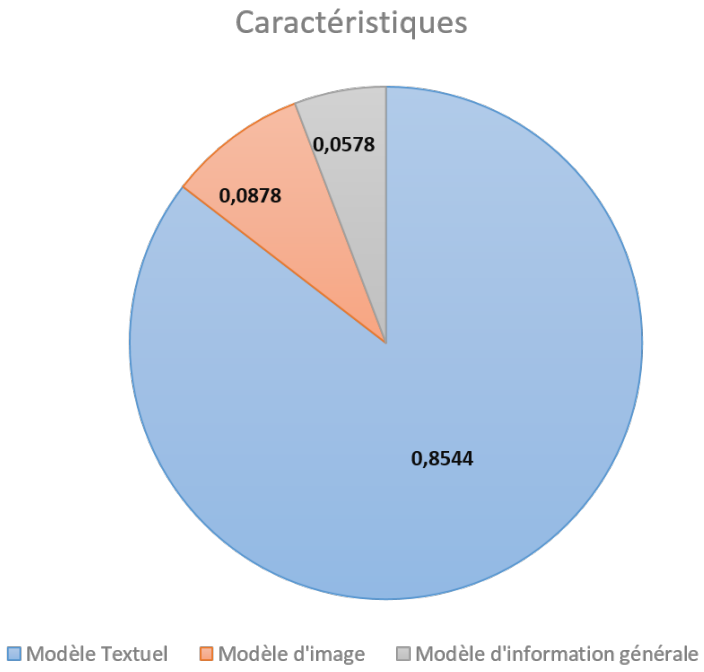

Figure 7.7: Importance des fonctionnalités de notre modèle : estimation des poids.

Comme nous l'avons mentionné ci-dessus, notre modèle s'adapte aux changements dans le temps. C'est pourquoi nous avons implémenté une composante qui ré-entraîne et rétablit un nouveau modèle. Pour cela, nous avons créé une base de données dans laquelle nous avons stocké le score du dernier modèle formé. Nous avons utilisé une fonction Python, qui vérifie à chaque changement de modèle si le score s'est amélioré après avoir ré-entraîné le modèle sur les données d'un nouvel utilisateur terroriste.

7.3.3 Évaluation des performances de notre modèle et discussion

L'étude expérimentale réalisée a permis de comparer les résultats obtenus avec ceux d'approches similaires de la littérature en utilisant notre jeu de données de test. Le tableau 7.5 présente un résumé des performances des approches. Par conséquent, la qualité des résultats de notre framework par rapport aux autres méthodes est parfaitement claire. Une évaluation de notre travail, suivant les quatre critères d'évaluation présentés dans le chapitre 4, peut nous conduire à des interprétations pertinentes.

Pour la multimodalité des données (MD), nous pouvons dire que notre solution fonctionne très bien avec les trois types de données d'entrée (texte, image et informations générales) avec un score de précision de 98,788%, 94,665% et 91,889% respectivement. Par conséquent, elle dépasse les autres approches pour chaque type de données. Au niveau de la multidimensionnalité des réseaux (MN), notre framework prouve également son efficacité avec 95,144% de précision contre 89,457% pour la seule méthode [120], à notre connaissance, qui utilise des réseaux multidimensionnels. Pour la capacité de détection des communautés (CD), notre méthode assure également un bon résultat avec une précision de 96,148% par rapport à [118] qui n'enregistre qu'une précision de 91,115%. L'adaptation de notre approche au dynamisme du comportement des terroristes (DB) se distingue par sa performance avec 92.322% par rapport aux autres méthodes dont la précision se limite à 88%.

Suite aux résultats obtenus, nous pouvons déduire trois conclusions principales :

— L'utilisation de plusieurs types de données collectées à partir de divers réseaux sociaux permet d'améliorer le niveau de connaissance des personnes. L'enrichissement de ces données permet donc de compléter les données dont nous disposons déjà et de rapprocher les données existantes dans les réseaux sociaux. L'enrichissement des données est essentiel pour maintenir

Table 7.5: Résumé des performances des approches

Méthode	MD			MN	CD	DB
	TX	IM	GI			
[28]	96.018%	-	91.441%	-	-	-
[120]	-	-	-	89.457%	-	84.990%
[118]	93.562%	-	-	-	91.115%	87.714%
Notre framework	98.788%	94.665%	91.889%	95.144%	96.148%	92.322%

une bonne qualité des données. Il permet de disposer de données complètes, qualifiées et actualisées, ce qui probablement, permet d'augmenter la performance d'un tel processus de détection du terrorisme.

— Les techniques basées sur le Deep Learning constituent un moyen efficace d'analyser des réseaux sociaux. Les algorithmes de Deep Learning dépassent les capacités humaines lorsqu'il s'agit de classer tout type de données et s'adaptent très bien aux changements dans le temps.

— La détection de comportements anormaux, et plus particulièrement des comportements terroristes, dans les réseaux sociaux en ligne est une tâche complexe à réaliser. Aucun signe n'est convaincant et cohérent pour juger qu'une personne est un terroriste dans ce type de réseau. Ceci est dû aux difficultés de collecter des données détaillées et privées sur les individus, étant donné la confidentialité appliquée dans les environnements virtuels en ligne.

7.4 Conclusion

Depuis 1979, les attaques organisées dans le monde via Internet ont tué au moins 436k personnes. Des moyens efficients doivent être mis en place par les gouvernements pour lutter contre le terrorisme. En outre, il est nécessaire d'identifier

soigneusement les groupes terroristes et de comprendre leurs tactiques afin de pouvoir détecter et prévoir les personnes impliquées dans l'organisation de ces événements violents. Dans ce contexte, l'objectif principal de ce chapitre est de proposer un framework intelligent comme solution pour la détection de personnes terroristes sur plusieurs réseaux sociaux en même temps. Notre framework est divisé en deux modules principaux. Un premier module qui se charge de l'examen de la structure topologique d'un réseau multidimensionnel pour détecter les différentes communautés de personnes et ensuite calculer un score d'anomalie à partir d'une fonction, qui permet d'attribuer des scores bas pour les personnes atypiques et des scores élevés pour les personnes normales. Un modèle probabiliste de la distribution Beta a été utilisé afin d'identifier automatiquement les anomalies par la classification des différents scores estimés. Les profils des personnes identifiées comme anormales sont transmis au deuxième module de notre framework pour s'assurer qu'ils correspondent bien à des terroristes. Ce deuxième module exécute un modèle entraîné sur des données terroristes extraites de plusieurs sources hors-lignes et en ligne. Ce dernier se présente dans la sélection des caractéristiques publiques des profils et dans la transmission de ces caractéristiques à trois sous-modèles. Ensuite, une fonction de calcul est déclenchée afin de générer un score pour chaque profil analysé en fonction du poids de chaque sous-modèle, pour être ensuite classifié par un seuil de décision calculé automatiquement. Les résultats expérimentaux ont prouvé l'efficacité de notre Framework avec une valeur de précision d'environ 95%. Enfin, une extension de ce travail pourrait consister à élaborer un modèle de prédiction des personnes utilisant les réseaux sociaux susceptibles de devenir des terroristes, par l'analyse de leurs comportements sur une structure multidimensionnelle. Le travail présenté au sein de ce chapitre a été soumis au International Journal "Expert Systems with Applications" [13].

13. https ://www.journals.elsevier.com/expert-systems-with-applications

Chapitre 8

Conclusion et travaux futurs

Dans ce dernier chapitre, nous présentons d'abord une synthèse de nos travaux relatifs à la détection d'utilisateurs violents dans les réseaux sociaux en ligne. Nous introduisons également les limites et les perspectives à court et à long terme.

8.1 Synthèse

Les réseaux sociaux en ligne font partie intégrante de l'activité sociale quotidienne des gens. Ils fournissent des plateformes permettant de mettre en relation des personnes du monde entier et de partager leurs intérêts. Cependant, ces services de réseau ont également eu de nombreux impacts négatifs. en effet, l'existence de phénomènes d'agressivité et d'intimidation dans ces espaces est inévitable et doit donc être abordée. L'exploration des données des réseaux sociaux pour détecter les comportements violents et les menaces est un défi pour plusieurs raisons pratiques. Premièrement, des personnes différentes ont des façons différentes d'exprimer le même comportement violent. Il est difficile d'entraîner un modèle qui fonctionne pour tout le monde en raison de la variété des comportements et des diverses manières dont ils sont exprimés, comme des images, des vidéos et des commentaires exprimés dans différentes langues. Deuxièmement, la multimodalité et l'ultra-haute dimensionnalité des données structurées et non structurées disponibles sur les sites de réseaux sociaux rendent difficile le développement de méthodes d'exploration

de données capables d'extraire des informations pertinentes utiles à la détection de comportements violents. Enfin, les algorithmes doivent tenir compte de la nature variable dans le temps des réseaux pour traiter les nouveaux utilisateurs et liens et mettre automatiquement à jour les modèles construits.

Compte tenu des défis mentionnés ci-dessus, l'objectif de cette thèse est de construire des approches avancées pour l'extraction de données complexes de réseaux sociaux pour l'analyse prédictive. Par conséquent, ce travail de recherche est conçu pour développer des moyens efficaces pour l'exploration de données de réseaux sociaux, avec un accent particulier sur la découverte de connaissances de haut niveau telles que les modèles et les structures et leur utilisation dans la reconnaissance et la prédiction de comportements violents. Pour atteindre l'objectif susmentionné, les contributions suivantes sont proposées dans le travail de cette thèse :

1. La proposition d'un modèle d'exploration de la structure topologique d'un réseau monodimensionnel afin d'identifier et prédire les comportements violents. Ce modèle a été entrainé sur des données terroristes de divers types extraites de Twitter. Les résultats expérimentaux ont marqué une performance remarquable par rapport aux travaux existants de la lilttérature, mais ils ne sont pas très optimaux.
2. Le développement d'une méthode non supervisée de détection des noeuds atypiques dans une structure multidimensionnelle basée sur l'analyse des relations entre les différents noeuds du réseau afin d'extraire leurs proporiétés graphiques. Ces propriétés sont utilisées pour déterminer le regrouppement communautaire des noeuds ainsi que le degré d'influence de chaque noeud dans sa communauté. Les expérimentations ont montré de bons résultas dans le contexte de détection des noeuds à comportement anormal.
3. La construction d'un framework qui traite à la fois l'aspect comportemental

et structurel des réseaux sociaux afin d'étendre les deux contributions précédentes. Ce framework a amélioré considérablement les résultats précédents en proposant une solution manipulant une structure multidimensionnelle dynamique analysant plus de types de données par rapport à la première contribution, et explorant les relations et les activités des utilisateurs par rapport à la deuxième contribution. Les résultats obtenus sont très performants, ils ont atteint une précision supérieure à 95%.

8.2 Travaux futurs

Les résultats satisfaisants obtenus de notre framework ouvrent de nouvelles perspectives de recherche. En effet, la solution proposée pourra être étendue selon plusieurs directions :

— Comme première extension immédiate, nous avons l'intention d'appliquer les fonctionnalités et les critères mis en valeur dans les contributions de cette thèse dans d'autres domaines liés à l'évaluation des sources tels que la fraude, le spam et les rumeurs. Nous essaierons de générer des réseaux synthétiques avec des caractéristiques du monde réel afin d'étudier leur évolution et leur influence. La prédiction et la recommandation de liens sont des pistes de recherche prometteuses dans le domaine de l'intégration de graphes.

— Une deuxième extension à long terme consisterait à cibler les groupes plutôt que les individus en détectant des changements dans l'évolution de l'activité d'un groupe et en déterminant si l'évolution actuelle est relativement conforme ou si elle s'écarte de l'évolution normale. Ce travail peut définir un nouvel axe de recherche pour la définition des concepts de base de l'évolution de l'activité des communautés en appliquant des caractéristiques

historiques.

Bibliographie

[1] R. Kaur and S. Singh. A survey of data mining and social network analysis based anomaly detection techniques. Egyptian Informatics Journal, 17(2) : 199–216, 2016.

[2] Inès Ben Kraiem. Détection d'anomalies multiples par apprentissage automatique de règles dans les séries temporelles. Thèse de doctorat, université de Toulouse, 2021. URL `https://hal.archives-ouvertes.fr/tel-03137163/document`.

[3] M. Fire, R. Goldschmidt, and Y. Elovici. Online social networks : threats and solutions. IEEE Communications Surveys & Tutorials, 16(4) :2019–2036, 2014.

[4] Simon Kemp. Digital 2021 : Global overview report. Technical report. available at `https://datareportal.com/reports/digital-2021-global-overview-report`.

[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks : Structure and dynamics. Physics reports, 424(4) :175–308, 2006.

[6] S. Strogatz. Exploring complex networks. Nature, 410 :268–276, 2001.

[7] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, and D. Pedreschi. Foundations of multidimensional network analysis. In 2011 International Conference on Advances in Social Networks Analysis and Mining, pages 485–489, 2011.

[8] P. Holme and J. Saramäki. Temporal networks. Physics Reports, 519(3) : 97–125, 2012.

[9] R. Hassanzadeh, R. Nayak, and D. Stebila. Analyzing the effectiveness of graph metrics for anomaly detection in online social networks. In Web Information Systems Engineering - WISE 2012, pages 624–630, 2012.

[10] S. Boccaletti, G. Bianconi, R. Criado, C. D. Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. Physics Reports, 544 :1–122, 2014.

[11] A. Qazi and M. S. Khan. Exploring probabilistic network-based modeling of multidimensional factors associated with country risk. Risk analysis : an official publication of the Society for Risk Analysis, 41(6) :911–928, 2020.

[12] H. Li, J. Jing, H. Fan, Y. Li, Y. Liu, and J. Ren. Identifying cultural heritage corridors for preservation through multidimensional network connectivity analysis — a case study of the ancient tea-horse road in simao, china. Landscape research, 46(1) :96–115, 2021.

[13] R. Jingjing, D. J. Dubois, D. Choffnes, A. M. Mandalari, R. Kolcun, and H. Haddadi. Information exposure from consumer iot devices : A multidimensional, network-informed measurement approach. In Proceedings of the Internet Measurement Conference, pages 267–279, 2019.

[14] Z. Xiaokang, L. Wei, W. K. I-Kai, H. Runhe, and J. Qun. Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data. IEEE Transactions on Emerging Topics in Computing, 9(1) :246–257, 2021.

[15] T. Tian, Q. You, L. Zhang, X. Yi, H. Yan, W. Xu, and Z. Su. Sorghumfdb : sorghum functional genomics database with multidimensional network analysis. Database (Oxford) : : the journal of biological databases and curation, 2016 :1–16, 2016.

[16] N. Papachristou, P. Barnaghi, B. Cooper, K. M. Kober, R. Maguire, S. M. Paul, M. Hammer, F. Wright, J. Armes, E. P . Furlong, L. McCann, Y. P . Conley, E. Patiraki, S. Katsaragakis, J. D. Levine, and C. Miaskowski. Network analysis of the multidimensional symptom experience of oncology. Scientific Reports, 9(1), 2019.

[17] G. Wang, Y. Wang, J. Li, and K. Liu. A multidimensional network link prediction algorithm and its application for predicting social relationships. Journal of Computational Science, 53 :101–358, 2021.

[18] F. E. Grubbs. Procedures for detecting outlying observations in samples. Technometrics, 11(1) :1–21, 1969.

[19] Q. Ding, N. Katenka, P. Barford, E. Kolaczyk, and M. Crovella. Intrusion as (anti)social communication : Characterization and detection. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 886–894, 2012.

[20] L. Amic, V. Delage, V. Denise, A. Flambert, M. Hamel, K. Hamilton, S. Moriniere, D. Reynie, and M.Tchounikine. Les attentats islamistes dans le monde 1979-2019. Technical report, 2019. available at `https://www.fondapol.org/etude/les-attentats-islamistes-dans-le-monde-1979-2019/`.

[21] G. Weimann. Terrorist migration to social media. Georgetown Journal of International Affairs, 16(1) :180–187, 2015.

[22] M. Fire, G. Katz, and Y. Elovici. Strangers intrusion detection - detecting spammers and fake profiles in social networks based on topology anomalies. ASE Human Journal, 1(1) :26–39, 2012.

[23] E. Zheleva, L. Getoor, J. Golbeck, and U. Kuter. Using friendship ties and family circles for link prediction. In International Workshop on Social Network Mining and Analysis (SNAKDD), Advances in Social Network Mining and Analysis, pages 97–113, 2008.

[24] L. Akoglu, M. McGlohon, and C. Faloutsos. Oddball : spotting anomalies in weighted graphs. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Advances in Knowledge Discovery and Data Mining, volume 13, pages 410–421, 2010.

[25] A. Rezaei, Z. M. Kasirun, V. A. Rohani, and T. Khodadadi. Anomaly detection in online social networks using structure based technique. In Eighth International Conference on Internet Technology and Secured Transactions (ICITST), pages 619–622, 2013.

[26] Z. Chen, W. Hendrix, and N. F. Samatova. Community-based anomaly detection in evolutionary networks. Journal of Intelligent Information Systems, 39 :59–85, 2012.

[27] A. Chouchane and M. Bouguessa. Identifying anomalous nodes in multidimensional networks. In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 601–610, 2017.

[28] H. Alvari, S. Sarkar, and P. Shakarian. Detection of violent extremists in social media. In 2019 2nd International Conference on Data Intelligence and Security (ICDIS), pages 43–47, 2019.

[29] P. Chitrakar, C. Zhang, G. Warner, and X. Liao. Social media image retrieval using distilled convolutional neural network for suspicious e-crime and terrorist account detection. In 2016 IEEE International Symposium on Multimedia (ISM), pages 493–498, 2016.

[30] N.E.H Ben Chaabene, A. Bouzeghoub, R. Guetari, S. Balti, and H. Hajjami Ben Ghezala. Detection of users' abnormal behavior on social networks. In International Conference on Advanced Information Networking and Applications (AINA), Advanced Information Networking and Applications, volume 1151, pages 617–629, 2020.

[31] P. Choudhary and U. Singh. Terrorist migration to social media. International Journal of Computer Applications, 112(9) :24–29, 2015.

[32] M. Netzer, K. G. Kugler, L. A. Müller, K. M. Weinberger, A. Graber, C. Baumgartner, and M. Dehmer. A network-based feature selection approach to identify metabolic signatures in disease. Journal of Theoretical Biology, 310 :216–222, 2012.

[33] P. Irofti, A. Pătras, and A. Băltoiu. Quick survey of graph-based fraud detection methods. Technical report, 2021. available at `https://arxiv.org/abs/1910.11299v3`.

[34] S. A. Kokatnoor and B. Krishnan. Self-supervised learning based anomaly detection in online social media. International Journal of Intelligent Engineering and Systems, 13(3) :446–456, 2020.

[35] D. Chen, Q. Zhang, G. Chen, C. Fan, and Q. Gao. Forum user profiling by incorporating user behavior and social network connections. In International Conference on Cognitive Computing (ICCC), pages 30–42, 2018.

[36] Z. Zamanian, A. Feizollah, N. B. Anuar, L. B. M. Kiah, K. Srikanth, and S. Kumar. User profiling in anomaly detection of authorization logs. In Computational Science and Technology, volume 481, pages 59–65, 2019.

[37] A. H. Lashkari, M. Chen, and A. A. Ghorbani. A survey on user profiling

model for anomaly detection in cyberspace. Journal of Cyber Security and Mobility, 8 :75–112, 2019.

[38] M. Jaouadi and L. Ben Romdhane. Influence maximization problem in social networks : An overview. In 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), pages 1–8, 2019.

[39] M. E. Rakoczy, A. Bouzeghoub, A. L. Gancarski, and K. Wegrzyn-Wolska. Reputation prediction using influence conversion. In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), pages 43–48, 2018.

[40] M. E. Rakoczy, K. Wegrzyn-Wolska A. Bouzeghoub, and A. L. Gancarski. Exploring interactions in social networks for influence discovery. In International Conference on Business Information Systems (BIS), pages 23–37, 2019.

[41] V. Albu. Measuring customer behavior with deep convolutional neural networks. Broad Research in Artificial Intelligence and Neuroscience, 7(1) : 74–79, 2016.

[42] C. L. Z. Zhang, P. Luo, and X. Tang. Learning social relation traits from faceimages. In International Conference on Computer Vision (ICCV), pages 3631–3639, 2015.

[43] N. Banovic, T. Buzali, F. Chevalier, J. Mankoff, and A. K. Dey. Modeling and understanding human routine behavior. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 248—-260, 2016.

[44] Mohamed Daddi Hammou. Analyse du comportement du consommateur dans le marché algérien des assurances. Mémoire ingénieur, ENSSEA Alger, 2010.

[45] M. E. Rakoczy, A. Bouzeghoub, K. Wegrzyn-Wolska, and A. G. Lopes. Users views on others–analysis of confused relation-based terms in social network. In OTM Confederated International Conferences On the Move to Meaningful Internet Systems, pages 155—-174, 2016.

[46] F. Ramiandrisoa and J. Mothe. Profil utilisateur dans les réseaux sociaux : Etat de l'art. In CORIA 2017 - Conférence en Recherche d'Informations et Applications - 14th French Information Retrieval Conference, pages 395––404, 2017.

[47] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. Journal of the American Society for Information Science, 27(3) :129–146, 1976.

[48] J. Trajkova and S. Gauch. Improving ontology-based user profiles. In RIAO'04 : Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, pages 380—-390, 2004.

[49] D. Tchuente, M. F. Canut, N. Jessel, A. Péninou, and F. Sèdes. A community based algorithm for deriving users' profiles from egocentrics networks : experiment on facebook and dblp. Social Network Analysis and Mining, 3 (3) :667–683, 2013.

[50] T. S. Priya, P. T. Revathy, T. Pradeesh, and C. R. R. Robin. Design and development of an ontology based personal web search engine. Procedia Technology, 6 :299–306, 2012.

[51] N. Hernandez, J. Mothe, C. Chrisment, and D. Egret. Modeling context through domain ontologies. Information Retrieval, 10(2) :143–172, 2007.

[52] S. Sosnovsky and D. Dicheva. Ontological technologies for user modelling. International Journal of Metadata, 5(1) :32–71, 2010.

[53] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos. Understanding user behavior in online social networks : a survey. IEEE Communications Magazine, 51(9) :144–150, 2013.

[54] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger. Understanding online social network usage from a network perspective. In Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, pages 35—48, 2009.

[55] N. Boob and D. M. Dakhane. Mining usage profiles using fuzzy clustering and its applications. International Journal of Emerging Technology and Advanced Engineering, 2(2) :120–123, 2012.

[56] Vic Barnett and Toby Lewis. Outliers in Statistical Data. Wiley, 3rd edition, 1994.

[57] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In Proceedings of the 2001 ACM SIGMOD international conference on Management of data, pages 37–46, 2001.

[58] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection : A survey. ACM computing surveys (CSUR), 41(3) :1–58, 2009.

[59] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang. Anomaly detection in online social networks. Social networks, 39 :62–70, 2014.

[60] J. Gao, N. Du, W. Fan, D. Turaga, S. Parthasarathy, and J. Han. A multi-graph spectral framework for mining multi-source anomalies. In Graph embedding for pattern analysis, pages 205–227, 2013.

[61] A. Al Hasib. Threats of online social networks. International Journal of Computer Science and Network Security, 9(11) :288–293, 2009.

[62] T. Amin, O. Okhiria, J. Lu, and J. An. Facebook : A comprehensive analysis of phishing on a social system. Technical report, 2010. available at `https://courses.ece.ubc.ca/412/term_project/reports/2010/facebook.pdf`.

[63] V. B. Livshits and W. Cui. Spectator : Detection and containment of javascript worms. In USENIX Annual Technical Conference, pages 335–348, 2008.

[64] NASAA. Informed investor advisory : Social networking. Technical report, 2011. available at `http://www.nasaa.org/5568/informed-investor-advisory-socialnetworking/`.

[65] R. Lundeen, J. Ou, and T. Rhodes. New ways i'm going to hack your web app. Technical report, 2011. available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.452.2038&rep=rep1&type=pdf`.

[66] B. Krishnamurthy and C. E. Wills. On the leakage of personally identifiable information via online social networks. In Proceedings of the 2nd ACM Workshop on Online Social Networks, page 7–12, 2009.

[67] J. R. Douceur. The sybil attack. In International Workshop on Peer-to-Peer Systems (IPTPS), Peer-to-Peer Systems, volume 2429, pages 251–260, 2002.

[68] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and

characterizing social spam campaigns. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pages 35–47, 2010.

[69] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao. Follow the green : Growth and dynamics in twitter follower markets. In Proceedings of the 2013 Conference on Internet Measurement Conference (IMC), pages 163–176, 2013.

[70] C. Taylor. Startup claims 80% of its facebook ad clicks are coming from bots. Technical report, 2012. available at `http://techcrunch.com/2012/07/30/startup-claims-80-of-its-facebook-ad-clicks-are-coming-from-bots/`.

[71] H. Mao, X. Shuai, and A. Kapadia. Loose tweets : an analysis of privacy leaks on twitter. In Proceedings of the 10th annual ACM workshop on Privacy in the electronic society, pages 1–12, 2011.

[72] S. Torabi and K. Beznosov. Privacy aspects of health related information sharing in online social networks. In 2013 {USENIX} Workshop on Health Information Technologies (HealthTech), 2013.

[73] L. Scism and M. Maremont. Insurers test data profiles to identify risky clients. Technical report, 2010. available at `http://online.wsj.com/news/articles/SB10001424052748704648604575620750998072986`.

[74] J. Vicknair, D. Elkersh, K. Yancey, and M. C. Budden. The use of social networking websites as a recruiting tool for employers. American Journal of Business Education (AJBE), 3(11) :7–12, 2010.

[75] L. Humphreys. Mobile social networks and social practice : A case study of dodgeball. Journal of Computer-Mediated Communication, 13(1) :341–360, 2007.

[76] G. Friedland and R. Sommer. Cybercasing the joint : On the privacy implications of geo-tagging. In Proceedings of the 5th USENIX Conference on Hot Topics in Security (HotSec), pages 1–8, 2010.

[77] M. Deans. The story of amanda todd. Technical report, 2012. available at `http://www.newyorker.com/online/blogs/culture/2012/10/amanda-todd-michael-brutsch-and-free-speech-online.html`.

[78] J. Wolak, D. Finkelhor, K. Mitchell, and M. Ybarra. Online "predators" and their victims : Myths, realities, and implications for prevention and treatment. Psychology of Violence, 1 :13–35, 2010.

[79] Priya Pedamkar. Text mining vs natural language processing. Technical report, 2019. available at `https://www.educba.com/important-text-mining-vs-natural-language-processing/`.

[80] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv :1301.3781, 2013.

[81] Shivangi Singhal. Data representation in nlp. Technical report, 2019. available at `https://medium.com/@shiivangii/data-representation-in-nlp-7bb6a771599a`.

[82] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25 :1097–1105, 2012.

[83] E. Kauderer-Abrams. Quantifying translation-invariance in convolutional neural networks. arXiv preprint arXiv :1801.01450, 2017.

[84] Jürgen Schmidhuber. Deep learning in neural networks : An overview. Neural networks, 61 :85–117, 2015.

[85] Ben Cole. Transfer learning. Technical report, 2018. available at `https://searchcio.techtarget.com/definition/transfer-learning`.

[86] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection : Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5) :1285–1298, 2016.

[87] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. Journal of Big Data, 6(1) :1–48, 2019.

[88] L. C. Freeman. Centrality in social networks conceptual clarification. Social networks, 1(3) :215–239, 1978.

[89] S. Agarwal and A. Sureka. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In International Conference on Distributed Computing and Internet Technology, pages 431–442, 2015.

[90] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. Machine learning, neural and statistical classification. 1994.

[91] J. AK. Suykens and J. Vandewalle. Least squares support vector machine classifiers. Neural processing letters, 9(3) :293–300, 1999.

[92] R. E. Wright. Logistic regression. In Reading and understanding multivariate statistics, pages 217–244, 1995.

[93] L. D. Harmon. Artificial neuron. Science Journal, 129(3354) :962–963, 1959.

[94] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari. Content-based filtering in on-line social networks. In International Workshop on Privacy and Security Issues in Data Mining and Machine Learning (PSDML), pages 127–140, 2011.

[95] P. W. Holland and S. Leinhardt. The structural implications of measurement error in sociometry. Journal of Mathematical Sociology, 3(1) :85–111, 1973.

[96] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In 23rd {USENIX} Security Symposium, pages 223–238, 2014.

[97] C. Xiao, D. M. Freeman, and T. Hwa. Detecting clusters of fake accounts in online social networks. In Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, pages 91–101, 2015.

[98] L. Shu, H. Xu, and B. Liu. Doc : Deep open classification of text documents. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2901–2906, 2017.

[99] Y. Zhang, W. Chen, C. K. Yeo, C.T Lau, and B. S. Lee. Detecting rumors on online social networks using multi-layer autoencoder. In Proceedings of the 2017 IEEE Technology Engineering Management Conference (TEMSCON), pages 437—441, 2017.

[100] B. Yang, J. Cao, R. Ni, and L. Zou. Anomaly detection in moving crowds through spatiotemporal autoencoding and additional attention. Advances in Multimedia, 2018 :1–8, 2018.

[101] J. Glasser and B. Lindauer. Bridging the gap : A pragmatic approach to generating insider threat data. In 2013 IEEE Security and Privacy Workshops (SPW), pages 98–104, 2013.

[102] Z. Chen and B. Liu. Mining topics in documents : standing on the shoulders of big data. In Proceedings of the 20th ACM SIGKDD international

conference on Knowledge discovery and data mining (SIGKDD), pages 1116–1125, 2014.

[103] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. Science Journal, 359(6380) :1146–1151, 2018.

[104] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In Proceedings of the IEEE international conference on computer vision (ICCV), pages 2720–2727, 2013.

[105] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe. Abnormal event detection in videos using generative adversarial nets. In 2017 IEEE International Conference on Image Processing (ICIP), pages 1577–1581, 2017.

[106] J.M. Berger and B. Strathearn. Who matters online : Measuring influence, evaluating content and countering violent extremism in online social networks. Technical report, 2013. available at `https://apo.org.au/node/33405`.

[107] J.M. Berger and J. Morgan. The isis twitter census : Defining and describing the population of isis supporters on twitter. Technical report, 2015. available at `https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf`.

[108] L. Getoor and C. P. Diehl. Link mining : a survey. ACM SIGKDD Explorations Newsletter, 7(2) :3–12, 2005.

[109] J. Castellini, V. Poggioni, and G. Sorbi. Fake twitter followers detection by denoising autoencoder. In Proceedings of the International Conference on Web Intelligence (WI), pages 195–202, 2017.

[110] Z. Li, D. Sun, R. Zhu, and Z. Lin. Detecting event-related changes in organizational networks using optimized neural network models. PloS one, 12 (11) :1–21, 2017.

[111] J. Shetty and J. Adibi. The enron email dataset database schema and brief statistical report. In Information Sciences Institute, 2004.

[112] S. Atran. John jay artis transnational terrorism database. In College of Criminal Justice, 2009.

[113] G. LaFree and L. Dugan. Introducing the global terrorism database. Terrorism and political violence, 19(2) :181–204, 2007.

[114] K. Leetaru and P. A. Schrodt. Gdelt : Global data on events, location, and tone. ISA Annual Convention, 2 :1–49, 2013.

[115] D. Byman and J. Shapiro. We shouldn't stop terrorists from tweeting. The Washington Post, 9, 2014.

[116] J. G. Jaspersen and G. Montibeller. On the learning patterns and adaptive behavior of terrorist organizations. European Journal of Operational Research, 282(1) :221–234, 2020.

[117] J. U. Siebert and D. Von Winterfeldt. Comparative analysis of terrorists' objectives hierarchies. Decision Analysis, 17(2) :97–114, 2020.

[118] S. Tutun, M. T. Khasawneh, and J. Zhuang. New framework that uses patterns and relations to understand terrorist behaviors. Expert Systems with Applications, 78 :358–375, 2017.

[119] B. Mahmood and M. Alanezi. Structural-spectral-based approach for anomaly detection in facebook network : Iraqi demonstrations case study.

International Journal of Computing and Digital Systems, 10(1) :343–351, 2021.

[120] G. Kalpakis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris. Identifying terrorism-related key actors in multidimensional social networks. In International Conference on Multimedia Modeling (MMM), MultiMedia Modeling, pages 93–105, 2019.

[121] S. D. Bhattacharjee, J. Yuan, Z. Jiaqi, and Y. Tan. Context-aware graph-based analysis for detecting anomalous activities. In 2017 IEEE International Conference on Multimedia and Expo (ICME), pages 1021–1026, 2017.

[122] Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition : Natops aircraft handling signals database. In 2011 IEEE International Conference on Automatic Face Gesture Recognition (FG), pages 500–506, 2011.

[123] M. K. Hayat, A. Daud, A. A. Alshdadi, A. Banjar, R. A. Abbasi, Y. Bao, and H. Dawood. Towards deep learning prospects : insights for social media analytics. IEEE Access, 7 :36958–36979, 2019.

[124] R. Chalapathy and S. Chawla. Deep learning for anomaly detection : A survey. arXiv preprint arXiv :1901.03407, 2019.

[125] Y. Yang, YC. Guo, and YN. Ma. Characterization of communities in online social network. In Proceedings of 2010 Cross-Strait Conference on Information Science and Technology (CSCIST), pages 600–605, 2010.

[126] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. Knowledge-based systems, 46 :109–132, 2013.

[127] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. Journal of machine learning research, 12 :2493–2537, 2011.

[128] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science Journal, 290(5500) : 2319–2323, 2000.

[129] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández. Syntactic n-grams as machine learning features for natural language processing. Expert Systems with Applications, 41(3) :853–860, 2014.

[130] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in python. Journal of Machine Learning Research, 12 : 2825–2830, 2011.

[131] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web, pages 285–295, 2001.

[132] A. Hagberg, D. Schult, and P. Swart. Networkx reference–release 2.6.2. Technical report, 2021. available at `https://networkx.org/documentation/stable/reference/index.html`.

[133] I. Sawadogo, L. Odongo, and I. Ly. Maximum likelihood estimation of the parameters of exponentiated generalized weibull based on progressive type ii censored data. Open Journal of Statistics, 7(6) :956–963, 2018.

[134] Danho Djrobie. Modèle de mélange et classification. Mémoire, université

Paris-Dauphine - PSL, 2016. URL `https://www.ceremade.dauphine.fr/~vigeral/Memoire2016Danho.pdf`.

[135] F. Couton, M. Danech, and M. Broniatousk. Application des mélanges de lois de probabilité à la reconnaissance de regime trafic routier. Technical report, 1996. available at `https://trid.trb.org/view/988611`.

[136] Z. Ma and A. Leijon. Bayesian estimation of beta mixture models with variational inference. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(11) :2160–2173, 2011.

[137] S. Boutemedjet, D. Ziou, and N. Bouguila. Model-based subspace clustering of non-gaussian data. Neurocomputing, 73(10-12) :1730–1739, 2010.

[138] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society : Series B (Methodological), 39(1) :1–22, 1977.

[139] G. Schwarz. Estimating the dimension of a model. The annals of statistics, pages 461–464, 1978.

[140] J. Weaver and P. Tarjan. Facebook linked data via the graph api. Semantic Web, 4(3) :245–250, 2013.

[141] Kevin Makice. Twitter API : Up and running : Learn how to build applications with the Twitter API. O'Reilly Media, Inc., 1st edition, 2009.

ECOLE
DOCTORALE

Titre : Détection d'utilisateurs violents et de menaces dans les réseaux sociaux

Mots clés : Détection d'anomalie, Analyse des réseaux sociaux complexes, Données multimodales, Réseau multidimensionnel, Détection de communautés

Résumé : Les réseaux sociaux en ligne ont été conçus pour faciliter et maintenir la socialisation entre les personnes. Cependant, certaines personnes ont commencé à utiliser ces réseaux à des fins dangereuses comme le harcèlement, le crime, la cyberintimidation, etc. Ainsi, un besoin énorme d'approches et de systèmes efficaces de détection des comportements anormaux préservant l'aspect dynamique des réseaux sociaux est devenu nécessaire. Néanmoins, l'analyse de données volumineuses, complexes et multimodales figurant dans la structure des réseaux sociaux est une tâche fastidieuse.
Pour répondre à cette problèmatique, nous visons dans cette thèse à (i) fournir une représentation multidimensionnelle des profils des utilisateurs provenant de plusieurs plateformes de réseaux sociaux, (ii) désigner le type du comportement d'un utilisateur en analysant les relations entre les différents individus présents sur la structure multidimensionnelle du réseau construit et (iii) décider la nature des comportements utilisateurs par l'application d'un modèle entrainé sur un comportement spécifique dangereux.
Afin d'aboutir au premier objectif, nous collectons des profils réels des trois réseaux sociaux à savoir, Facebook, Twitter et Instagram qui comprennent des données multimodales pour des personnalités publiques appartenant à divers domaines. Pour ce faire, nous nous appuyons sur l'utilisation de plusieurs techniques et outils d'extraction de données pertinentes en garantissant la synchronisation entre ces données. Pour atteindre le deuxième objectif, nous estimons le degré d'influence de chaque individu sur son entourage. Pour soutenir cette idée, nous proposons une méthode basée sur la détection des communautés des utilisateurs en intégrant le calcul du degré externe normalisé de chaque pair d'individu dans le réseau. Pour répondre au troisième objectif, nous combinons l'aspect comportemental des profils et l'aspect structurel assuré dans les relations entre les profils pour garantir une analyse compléte sur le réseau. Un framework est conçu à cet effet pour exploiter un modèle d'apprentissage manipulant des données terroristes extraitres de plusieurs sources.
Pour valider l'efficacité de notre travail, nous utilisons différents moyens de l'apprentissage automatique afin de tester nos approches sur des jeux de données réels.

Title : Detection of violent users and threats in social networks

Keywords : Anomaly detection, Complex social networks analysis, Multimodal data, Multidimensionnal network, Community detection

Abstract : Online social networks were designed to facilitate and maintain socialization between people. However, some people have started to use these networks for dangerous purposes like harassment, crime, cyberbullying etc. Thus, a huge need for effective approaches and systems for detecting abnormal behaviors preserving the dynamic aspect of social networks has become necessary. However, analyzing large, complex and multimodal data in the structure of social networks is a hard task.
To respond to this problematic, we aim in this thesis to (i) provide a multidimensional representation of user profiles coming from several social networking platforms, (ii) designate the type of behavior of a user by analyzing the relationships between the different individuals present on the multidimensional structure of the constructed network and (iii) decide the nature of user behavior by applying a model trained on a specific dangerous behavior.
In order to achieve the first objective, we collect real profiles of the three social networks namely, Facebook, Twitter and Instagram which include multimodal data for public figures belonging to various fields. To do this, we use several techniques and tools for extracting relevant data by ensuring synchronization between these data. To reach the second objective, we estimate the degree of influence of each individual on his entourage. To support this idea, we propose a method based on the detection of user communities by integrating the calculation of the normalized external degree of each peer of individuals in the network. To ripost to the third objective, we combine the behavioral aspect of the profiles and the structural aspect ensured in the relations between the profiles to guarantee a complete analysis on the network. A framework is designed for this purpose to exploit a learning model handling terrorist data extracted from several sources.
To validate the effectiveness of our work, we use various means of machine learning to test our approaches on real data sets.

Institut Polytechnique de Paris
91120 Palaiseau, France