République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université Larbi Tébessi - Tébessa

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Mémoire de fin d'étude

Pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et informatique

Filière : Informatique

Option : Systèmes d'Informations

Thème

Prédiction des cascades de l'information sur le réseau twitter

Réalisé par :

Djeddi Anis

Devant le jury :

Dr. Menassel Rafik	MCA	Université Larbi Tébessi	Président
Dr. Boualleg Yakoub	MAA	Université Larbi Tébessi	Examinateur
Dr. Nait-Hamoud. M.C	MCB	Université Larbi Tébessi	Encadreur

Date de soutenance

30/06/2022

Remerciement

Tout d'abord, je remercie Dieu le Tout-Puissant de m'avoir donné la force morale etphysique et de m'avoir permis d'accomplir ce travail.

Je tiens à remercier mon encadrant ***Dr. Nait-Hamoud .M.C****, pour m'avoir dirigédans ce travail. Je le remercie pour sa*

Disponibilité, son suivi, ses précieux conseils et son aide.

Je tiens également à exprimer ma gratitude à Monsieur le Président du jury, ***Menassel Rafik****, et à Monsieur examinateur* ***Boualleg Yakoub****.*

Merci à mon père, à ma mère, à mes sœurs, à mes frères, pour leur patience et leur soutien. Alors, à toute ma famille.

Ils ont toujours été la source de mon succès, pour cela je ne saurai jamais commentles remercier.

Je prie Dieu de me les garder le plus longtemps possible

Dédicace

A mes chers parents,

Que nulle dédicace ne puisse exprimer ce que je leurs dois, pour leur bienveillance, leur affection et leur soutien…Trésors de bonté, de générosité et de tendresse, en témoignage de mon profond amour et ma grande reconnaissance « Que Dieu vous garde ».

A mes chères et sœurs

En témoignage de mes sincères reconnaissances pour les efforts qu'ilsont consenti pour l'accomplissement de mes études. Je leur dédie ce modeste travail en témoignage de mon grand amour et ma gratitude infinie.

A tous mes amis

Pour leur aide et leur soutien moral durant l'élaboration du travail defin d'études.

A tous ma Famille

A tous ceux dont l'oubli du nom n'est guère celui du cœur…

« De l'union « si » avec « mais » naquit enfant nommé « jamais » »

« Il n'y a pas de « si »ni de « mais », il faut réussir »

Anis

Résumé

La diffusion de l'information dans les réseaux sociaux, communément appeléecascade de l'information, est un domaine qui a été beaucoup étudié dans la littérature. Ce phénomène survient lorsque des individus observant les signaux de leurs voisins directs se mettent à propager leurs contenus sans consensus au préalable. L'analyse des cascades de 'information permet, entre autres, aux entreprises de se forger une idée pour le lancement d'un éventuel produit ; et aux gouvernements d'exploiter ce phénomène dans les processus de vote, ou pour l'adoption d'une politique. Les recherches dans ce domaine ont ciblé la prédiction descascades de l'information basée sur les caractéristiques des réseaux, le contenu des échanges et les propriétés temporelles des changes. L'objective de ce projet est d'explorer la prédiction des cascades de 'information à partir d'une large base de tweets réels obtenus à partir du micro blog Twitter.

Mots clés : Cascade d'information , Prédiction , Twitter , Diffusion d'information.

Abstract

The dissemination of information in social networks, commonly referred to as the information cascade, is an area that has been extensively studied in the literature. This phenomenon occurs when individuals observing the signals of their direct neighbors begin to propagate their content without prior consensus. The analysis of information cascades allows, among other things, companies to form an idea for the launch of a possible product; and to

governments to exploit this phenomenon in the voting process, or for the adoption of a policy. Research in this area has targeted the prediction of information cascades based on the characteristics of networks, the content of exchanges and the temporal properties of exchanges. The objective of this project is to explore the prediction of information cascades from a large base of real tweets obtained from the Twitter microblog.

Keywords: Information cascade , Prediction , Twitter , Information dissemination.

ملخص

إن نشر المعلومات في الشبكات الاجتماعية، الذي يشار إليه عادة باسم تسلسل المعلومات، هو مجال تمت دراسته على نطاق واسع في الأدبيات. تحدث هذه الظاهرة عندما يبدأ الأفراد الذين يراقبون إشارات جيرانهم المباشرين في نشر محتواهم دون إجماع مسبق. يسمح تحليل تسلسلات المعلومات، من بين أشياء أخرى، للشركات بتكوين فكرة لإطلاق منتج محتمل؛ والحكومات لاستغلال هذه الظاهرة في عملية التصويت، أو لتبني سياسة. استهدفت الأبحاث في هذا المجال التنبؤ بتسلسل المعلومات بناءً على خصائص الشبكات ومحتوى التبادلات والخصائص الزمنية للتبادلات. الهدف من هذا المشروع هو استكشاف تنبؤ تسلسل المعلومات من قاعدة كبيرة من التغريدات الحقيقية التي تم الحصول عليها من مدونة تويتر الصغيرة.

الكلمات المفتاحية: تسلسل المعلومات، التنبؤ، تويتر، نشر المعلومات.

Table de matière

Liste des figures

Introduction générale

Introduction générale

Lorsque quelqu'un remarque les articles, tweets, etc. d'une autre personne et commence à les partager ou à les republier, on parle de cascade d'informations. Le processus continue jusqu'à ce que les gens cessent de le diffuser ou de le publier. Selon la plate-forme de médias sociaux, l'activité peut inclure le partage d'opinions, de messages ou de contenu multimédia (comme une photo ou une vidéo). La diffusion d'informations pourrait être utile pour façonner les attitudes des utilisateurs. L'analyse de la diffusion de l'information peut être utile aux entreprises lors du lancement de nouveaux produits. Il peut également être appliqué pour comprendre les marchés boursiers . Le gouvernement et les partis politiques l'utilisent pour diffuser une idéologie et obtenir un soutien lors des élections ou de l'introduction de nouvelles politiques.

Le problème de la cascade a également été décrit comme un problème de classification, où la fréquence de transmission d'un message peut dépendre du temps, du contenu du message par rapport à ses caractéristiques temporelles ou de la popularité des tweets basés sur le tweet d'origine. Les recherches antérieures se sont principalement concentrées sur les caractéristiques des réseaux sociaux, telles que leurs réseaux d'utilisateurs, leur contenu informationnel et leurs propriétés temporelles. Récemment, en plus des caractéristiques basées sur le contenu, les chercheurs ont commencé à étudier les caractéristiques basées sur l'utilisateur ainsi que les propriétés structurelles et temporelles.

Dans ce projet, nous proposons une méthode de détection des cascades d'informations sur les réseaux Twitter à l'aide de la relation suiveur-suivi basée sur l'approche de la marche aléatoire et des mesures de similarité. L'approche proposée offre une plus grande précision que les approches existantes. a commencé à examiner les propriétés structurelles et temporelles de la fonctionnalité basée sur l'utilisateur.

Introduction générale

Cette thèse est organisée comme suit :

Chapitre 1 : Réseaux sociaux : concept de base

Chapitre 2 : Diffusion de l'information dans les réseaux sociaux

Chapitre 3 : Etude de cas

Finalement nous conclurons avec une conclusion générale et les perspectives

Chapitre I

Réseaux Sociaux : concepts de base

I.1 Introduction

Cette dernière décennie, les réseaux sociaux ont attiré l'attention de beaucoup d'acteurs ; dont : les chercheurs, les entreprises, les politiciens et les institutions gouvernementales. Ce gain d'intérêts est dû aux applications de plus en plus nombreuses visées par les acteurs précités. En effet, l'énorme quantité de données échangées à travers ces réseaux est analysée aux fins de plusieurs applications telles que la propagande, l'intelligence économique, la défense, etc. Ce phénomène grandissant devient une partie importante de notre vie quotidienne, des millions de personnes échangent une quantité énorme de des données sur Twitter, Facebook, Instagram, etc. Ces services de micro-blogging gratuits, permettent aux gens de s'inscrire, de diffuser, de partager du contenu et de suivre d'autres membres. Les réseaux collaboratifs sont un autre type de réseaux sociaux où les membres collaborent pour atteindre des objectifs communs, des exemples de tels sites incluent DBLP, ResearchGate, etc.

I.2 Les réseaux sociaux

Depuis plus d'un siècle, l'expression « réseau social » est utilisée pour décrire des structures sociales constituées de nœuds liés par un ou plusieurs types spécifiques d'interdépendance tels que les valeurs, les visions, les idées, les échanges financiers, l'amitié, l'intérêt, les conflits ou le commerce.

De nos jours, il existe de nombreux réseau social tels que Facebook, Twitter, Instagram, Telegram, Snapchat, Wikipedia, Pinterest, TikTok, etc. En effet, avec l'avènement de ces derniers, une science à part a émergé pour valider les théories des sociologues d'une part, et pour titrer bénéfice de ces réseaux par lesquels des millions de personnes sont accessibles dans un temps très court.

Dans littérature, il y a abus de langage qui engendre une confusion entre deux concepts à savoir : les réseaux sociaux et les médias sociaux qui sont employés pour signifier la même chose. Ces deux notions sont différentes, car les médias sociaux englobent les réseaux sociaux. La figure I.1 illustre quelques types de médias sociaux. Ces derniers englobent les projets collaboratifs, les blogs, les micro-blogs, les communautés de partage de contenu, et les mondes virtuels. Il faut noter que les réseaux sociaux entrent dans la catégorie des micro-blogs dont la différence

avec les blogs qui sont des journaux d'idées publiés en ligne réside

dans le fait que ces derniers ne gardent aucune mise à jour des interactions entre les utilisateurs.

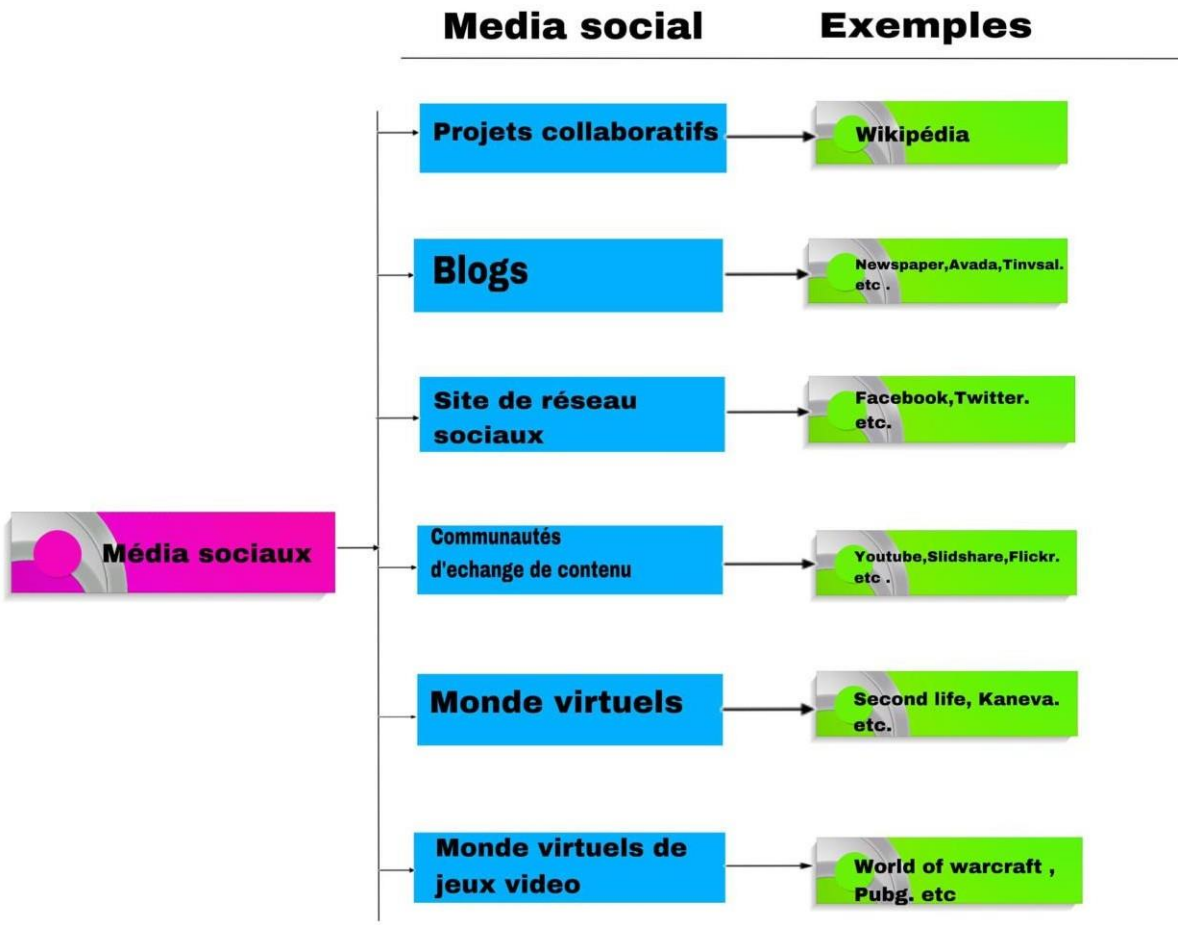

Figure I. 1:Quelques types de médias sociaux

I.1.1 Les propriétés des réseaux sociaux

Dans ce qui suit, nous introduisons q2uelques propriétés essentielles des réseaux sociaux qui ont émergé des études et des travaux expérimentaux dans plusieurs disciplines.

1) Six degrés de liberté

Cette propriété surprenante, a été étudiée dans les réseaux sociaux. En particulier, Twitter a été exploré pour enquêter sur la connectivité des utilisateurs. En effet, la société d'analyse de données Sysomos a découvert qu'en moyenne 50% des utilisateurs de Twitter sont à quatre pas les uns des autres ; alors que la distance qui sépare en moyenne deux utilisateurs, dans tout le réseau est de six pas. La figure I.2 montre les résultats de l'analyse de cette firme.

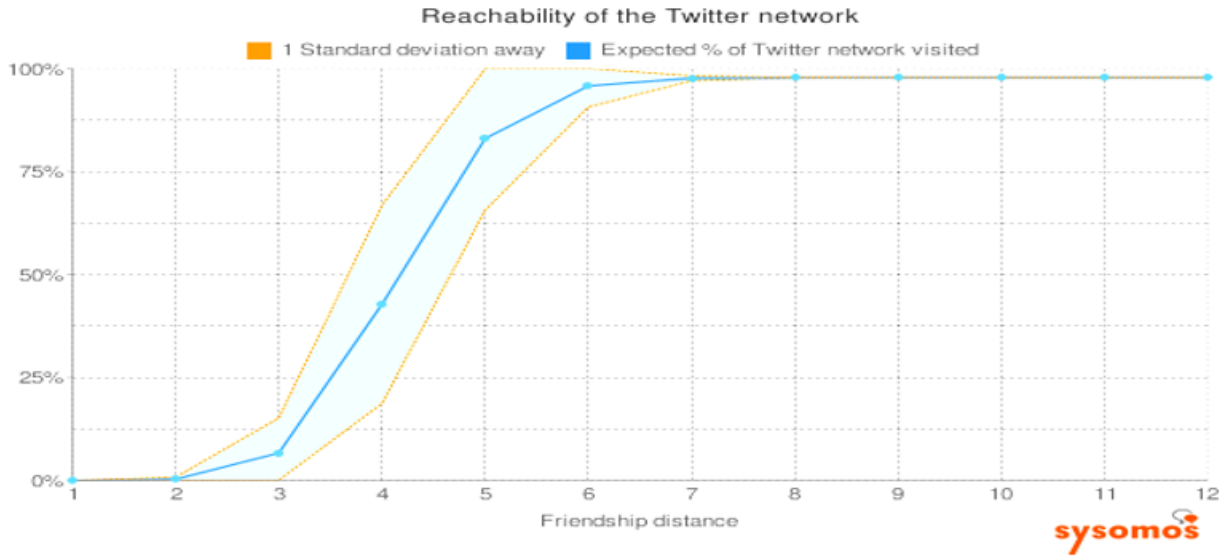

Figure I. 2:Illustration de la distance d'amitié sur Twitter [1]

2) Force des liens faibles

Il a été suggéré dans la littérature que les nouvelles informations dans les réseaux sociaux proviennent de liens faibles, c'est-à-dire, généralement les utilisateurs reçoivent des messages sur des thématiques nouvelles des voisins dont la cohésion avec le reste du groupe social est faible. Effectivement, ces résultats peuvent s'expliquer par le fait que les membres ayant de fortes relations dans un réseau social ont tendance à former des groupes dont les liens sont forts (en termes de relations inter utilisateurs) souvent appelés communautés. Par ailleurs, ces groupes sociaux partagent des intérêts communs et les mêmes thématiques. Par conséquent, il est plus probable que les nouvelles informations proviennent aux utilisateurs de membres extérieurs à la communauté ou situés à sa frontière. La figure I.3 ci-dessous illustre les liens forts et faibles en termes de relations inter utilisateurs.

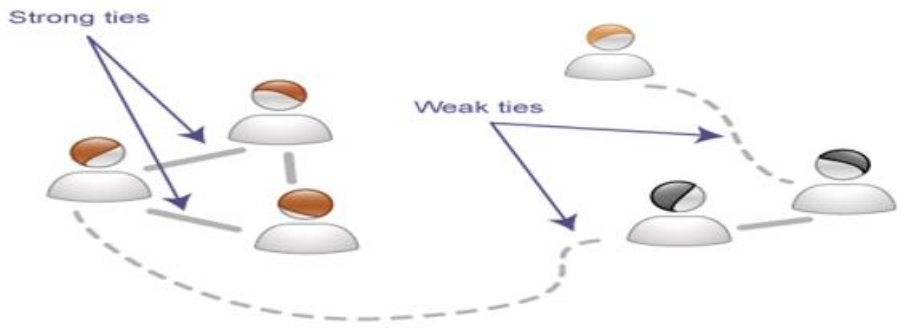

Figure I. 3:Force des liens faibles [2]

3) Fermeture triadique

Ce principe de transitivité des liens dans le réseau est appelé fermeture triadique, il s'exprime par le fait que le nombre de voisins communs entre deux nœuds augmente, c'est à dire, les amis des amis dans un réseau deviendront probablement des amis. En effet, si un nœud A est un ami du nœud B et que le nœud B est ami du nœud C, il est probable que les nœuds A et C deviendront amis dans un avenir proche. Ceci s'exprime part en termes de graphe que le triangle entre les nœuds A, B, et C est appelé à se refermer. La figure I.4 ci-dessous montre un exemple de la fermeture triadique qui se manifeste par la nouvelle relation entre Mohamed et Ali qui ont commun le même ami Anis.

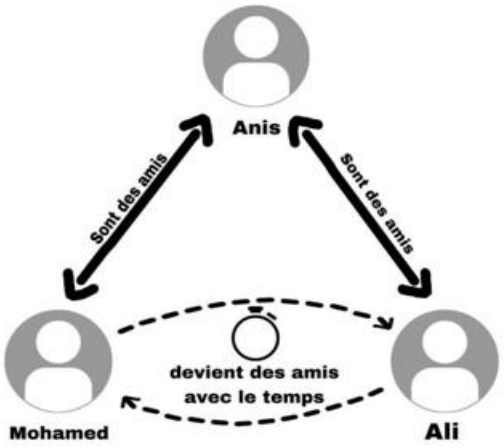

Figure I. 4:Fermeture triadique

4) L'homophilie

Dans les réseaux sociaux, l'homophilie signifie que les utilisateurs ont tendance à intéragir avec d'autres utilisateurs auxquels ils s'identifient de par leurs ressemblances selon des attributs tels que l'age, le sexe, l'origine ethnique, la langue, la profession, etc. Certaines études dans la léterature ont montré cette propriété. En effet, le chercheur De Choudhury [3] analysé dans des expérimentations 25.9 millions de tweets afin de montrer l'importance de l'homophilie dans la constitustion des liens dans le réseau social Twitter, afin de mesurer l'homophilie, l'auteur a utilisé plusieurs attributs dont les attributs démographique, ceux relatifs aux activités des utilisateurs sur Twitter, et les contenus des messages propagés entre les utilisateurs.

L'homophilie est une propriété très corrélée avec la fermeture triadique. En effet, ce constat peut être justifier par le fait que les utilisateurs qui ont des amis en communs se ressemblent.

5) L'influence sociale

Cette propriété se manifeste par le changement du comportement qu'un utilisateur d'un réseau social peut avoir à la suite de ses interactions avec ces voisins. En d'autres termes, c'est le changement de comportement, de cognition, d'attitude, de sentiment d'un utilisateur de réseau, qui émane de ses échanges avec ses voisins. La figure I.5 illustre cette propriété, les flèches en pointillés indiquent la relation entre des utilisateurs dans le graphe sous-jacent d'un réseau. Spécifiquement, dans le cas du réseau social Twitter, les flèches representent la relation *'Follow'*. Les flèches en bleu et rouge indiquent le flux d'information entre les utilisateurs. Cette figure illustre un exemple d'influence sur les réseaux sociaux, les messages postés par l'utilisateur U_1 ont été propagés dans le réseau par ses voisins, et les voisins de ses voisins. L'utilisateur U_8 a diffusé le contenu qu'il a reçu de U_1, cependant, il a ignoré celui posté par U_3. Ces comportements des utilisateurs sur les réseaux sociaux sont dûs à l'influence qu'exerce les utilisateurs les uns sur les autres. Des études sociologiques ont montré que les utilisateurs changent leurs comportements afin de ressembler à leurs amis. Le sociologue Kandel [4] differencie entre l'homophilie et l'influence social qu'il considère comme un processus où l'interaction entre les personnes engendre leurs ressemblances.

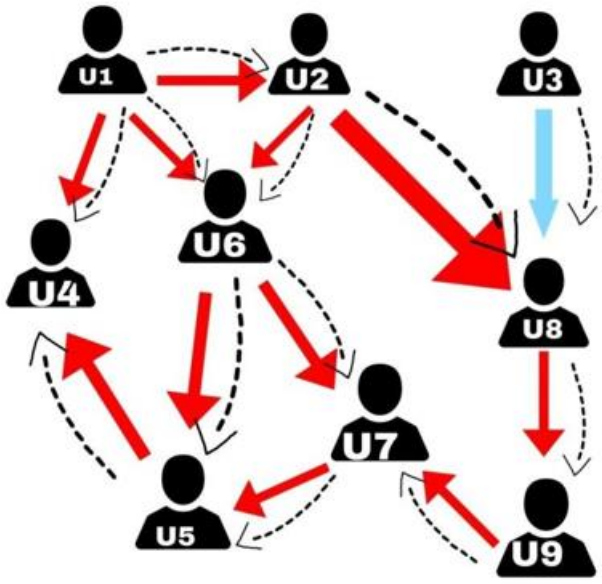

Figure I. 5:Illustration de l'influence sociale

I.1.2 Twitter

En mars 2006, un petit service de communication appelé Twitter a fait ses débuts. Cela a commencé comme un projet dans une société de podcasting de San Francisco. De nos jours, aprés quelques années de son apprution, Twitter est en plein essor. Ce réseau social est un service de micro-blogging gratuit, les utilisateurs peuvent s'abonner, partager instantanément du contenu et suivre d'autres membres. Le réseau social Twitter possède son propre jargon dont nous citons :

- Timeline : C'est une sorte pile de messages qui affiche le flux de tweets des membres Twitter suivis dans l'ordre chronologique de leur reception; l'utilisateur peut répondre à ces tweets, retweeter ou aimer un tweet depuis la chronologie.
- Post : Il désigne tout message publié sur Twitter pouvant contenir des photos, des vidéos, des liens. Contrairement à Facebook ou LinkedIn, les tweets sont consultables et publics, ils sont des messages courts envoyés sous forme de messages textes aux abonnés. Un tweet doit comporter moins de 140 caractères, les membres peuvent ajouter des hashtags ou des liens hypertexgte dans leurs posts.

- Hashtag : Il désigne un terme préfixé par le symbole dièse (#). Les hashtags sont utilisés pour signifier qu'un message concerne un sujet particulier et de faciliter la recherche du contenu. C'est un moyen d'indexation du text adopté par les utilisateurs

  par une convention définie par ces derniers en communs accords. Ce type de convention est appellé *Folksomonie*.
- Suivre : lorsque un utilisateur est intéressé par d'autres utilisateurs de Twitter, il vous s'abonne à leur mise à jour Twitter. Avec cette action d'abonnement, il sera abonné à eux, par conséquent, chaque fois que ces derniers publient un nouveau tweet, il apparaîtra sur sa timeline dans Twitter.
- Retweet : c'est le fait de poster un tweet d'une personne qu'un utilisateur suit sur le réseau et qu'il partage avec ses suiveurs (*followers)*.
- Message direct : c'est un message privé envoyé à utilisateur à un autre d'une manière directe sur Twitter.

Twitter devient de plus en plus utilisé, de nos jours, on recense plus de 6000 tweets publiés par seconde ce qui correspond à 500 millions de tweets postés par jour [5] . Les statistiques annocées par la societé BrandWatch montrent l'importance de ce réseau social nous citons

quelques chiffres ci-dessous:

- 330 millions d'utilisateurs sont actifs par mois

- Aux états unis d'amériques 22% de la population sont abonnés, et 10% des utilisateurs publient 80% des tweets.

- 24.6% des comptes Twitter sont détenus par des journalistes.

- 83% des leaders mondiaux sont abonnés à ce réseau.

-etc.

Ces chiffres justifient le phénomène et l'ampleur des réseaux sociaux dans la société contemporaine. Par conséquent, l'étude des réseaux sociaux qui est devenue une science à part constitue un moyen efficace d'extraction des connaissances à partir des contenus échangés par les utilisateurs. En effet, ces réseaux sont déjà analysés; particulièrement, le phénomène de diffusion de l'information qui suscite beaucoup d'intérêt. Ce dernier concerne l'étude de la propagation de l'information dans les réseaux sociaux.

I.3La diffusion de l'information

La diffusion de l'information est le processus qui étudie la propagation d'information dans les réseaux sociaux en ligne. Elle impacte la société vu la rapidité de l'adoption des réseau sociaux par un grand nombre d'abonnés. L'énorme quantité des contenus diffusés via ces réseaux attirent l'attention de plusieurs acteurs tels que les chercheurs scientifiques, les entreprises, et les agences gouvernementales. En effet, ces contenus sont analysés à des fins d'extraction de connaissances utiles à la prise de décision pour ces différents acteurs. La figure I.6 ci-dessous résume les différents types de diffusion de l'information qui représentent, de nos jours, un domaine à part entière. En effet, ces types différent selon l'obsevabilité du réseau et la disponibilité de l'information pour un utilisateur donné. L'observabilité du réseau signifie qu'il existe des relations entre les utilisateurs, géneralement, sous forme d'un graphe d'amitié ou de follower-followee dans le cas du réseau social Twitter. Si le réseau est observable et les utilisteurs perçoivent les informations d'une manière globale (i.e, toutes les informations partagées sur le réseau) alors la diffusion de l'information est dite *comportement grégaire*. Un exemple de ce comportement peut être observé sur les réseaux des enchères en lignes. Le réseau est explicite et les utilisateurs peut percevoir toutes les enchères de tous les utilisateurs. Dans le cas ou les utilisateurs ne pérçoivent que les informations qui proviennent de leurs voisins, la

diffusion de l'information s'appelle alors cascade de l'information. Ce type est généralement associés aux réseaux sociaux en ligne. Par ailleurs, lorsqu'il le réseau n'est pas observable, la diffusion de l'information s'apparente ou bien à la disffusion d'innovations ou à l'épédémie. Ces deux phénomènes sont similaires, sauf que dans le cas de l'innovation les idées, les produits ou les objets remplacent les pathogènes dans le cas de l'épédémie, et l'adoption remplace l'infection.

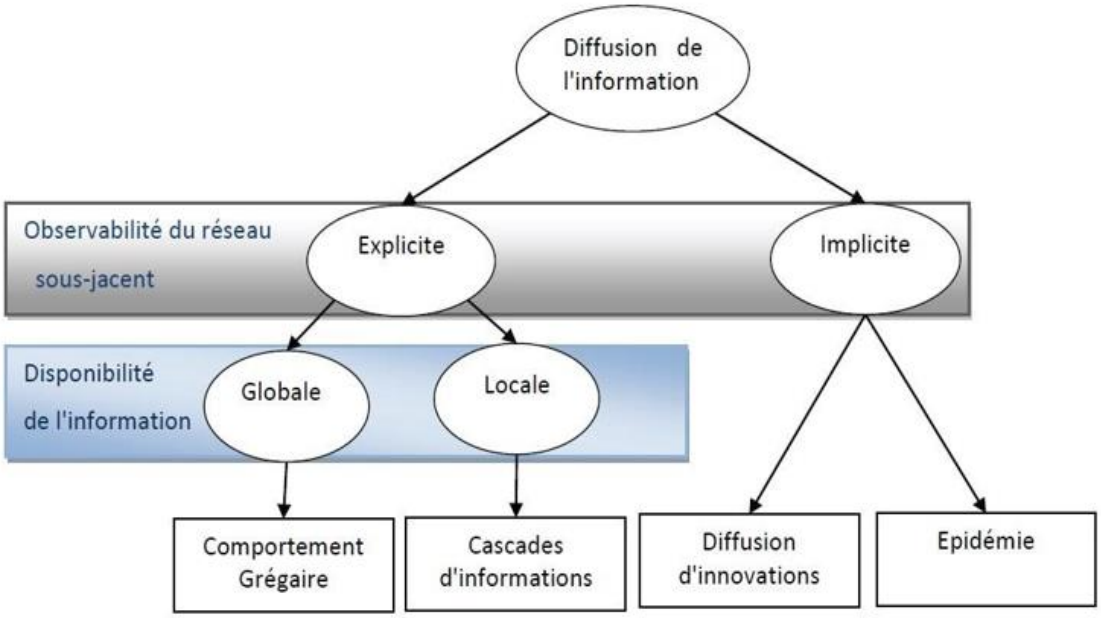

Figure I. 6:Types de diffusion de l'information [7]*Cascades d'information*

I.4Cascades d'information

Les cascades d'informations sont des processus dynamiques importants dans les réseaux complexes. Une cascade d'informations peut décrire la dynamique de diffusion de rumeurs, de mèmes que des campagnes de marketing, qui partent initialement d'un nœud ou d'un ensemble de nœuds du réseau. La figure I.6 illustre une cascade d'information désignée par la propagation de contenu C_1 à travers le réseau qui est indiquée par des fléches vertes sur la figure. Ce phénomene est important dans la mesure où il permet d'identifer les informations importantes qui se propagent sur le réseau. Aussi, l'identification des utilisateurs dont les messages se sont propagés sur le réseau offre une information pertinante pour un certain nombre d'acteurs. En effet, les entreprises de marketing s'intéressent à ces utilisateurs vu leurs aptitudes à influencer un grand nombre d'utilisateur du réseau, et la rapidité du processus de propagation des

informations émise par ces derniers. Par ailleurs, les agences gouvernementales de sécurité s'intéressent aussi à ce phénomène qui leur permet d'identifier les contenus interdits par les lois en vigueur, ainsi que les utilisateurs qui diffusent ces contenus.

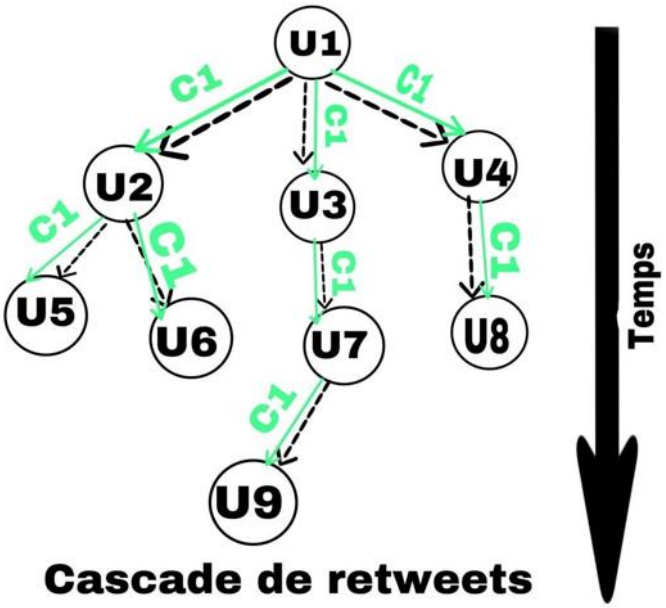

Figure I. 7:Cascades d'information

La diffusion de l'information sera présentée vec de plus amples détails, à travers l'étude de l'état de l'art, dans le prochain chapitre.

I.5 Conclusion

Les réseaux sociaux en ligne jouent un rôle important dans la société, ce chapitre a été consacré à leur introduction ainsi qu'à certains de leurs concepts fondamentaux. Egalement, nous avons mentionné certaines des propriétés intéressantes des réseaux sociaux qui ont émergé des études et des travaux expérimentaux, dans plusieurs disciplines. Par ailleurs, nous avons mis en évidence la notion de la diffusion de l'information dans les réseaux sociaux en ligne, particuliérement, celle des cascades des informations qui constituent l'objet de notre étude. Dans ce qui suit, nous utiliserons les appelation diffusion de l'information et cascades de l'information interchangeablement.

Chapitre II :

Diffusion de l’information dans les réseaux sociaux

II.1 Introduction

L'intérêt accordé à la diffusion de l'information a engendré beaucoup de travaux dans la littérature. Plusieurs aspects de cette problématique ont été traités dans le but de déterminer ce qui gouverne ce phénomène. Ceci inclus différentes tâches, telles que: la détection automatique des événements, la prédiction des cascades de l'information [48] et la détection des influenceurs . La détection des événements consiste à identifier les thématiques saillantes, c'est à dire, qui apparaissent dans le réseaux et déclinent rapidement. L'identification des influenceurs est aussi une application très recherchée, elle consiste à détecter les utilisateurs des réseaux sociaux dont les messages sont souvent propagés à travers le réseau par un grand nombre d'utilisateurs. Ces utilisateurs sont appelés influenceurs, car leurs suiveurs (followers) sur le réseau ont tendance à propager systématiquement leurs contenus. L'ex-président des états unis Barrack Obama est un influenceur sur Twitter. En effet, ce dernier a comptabilisé plus de onze (11) millions de suiveurs sur ce réseau. Ceci suggère que ces super utilisateurs ont acquis une certaine notoriété qui leur permet d'influer sur le comportement des utilisateurs, à un point où ces derniers ignorent leurs propres signaux et diffusent le signal de l'influenceur. Ces super utilisateurs sont de nos jours recrutés par des entreprises et par les gouvernements pour des fins commerciale, politique ou autres. La prédiction des cascades de l'information est aussi une problématique qui reste toujours d'actualité, elle consiste à prédire si un contenu sera propagé à un stade précoce de l'analyse du réseau en ligne.

Ce chapitre est consacré à la définition du problème étudié, et à la spécification des travaux antérieurs dans ce domaine.

Par ailleurs, afin de positionner notre problématique par rapport au domaine de la diffusion de l'information dans les réseaux sociaux, dans ce qui suit nous présentons une taxonomie des recherches de l'état de l'art.

II.2 Taxonomie des travaux sur la diffusion de l'information

La taxonomie des travaux de la littérature sur la problématique de la diffusion de l'information est illustrée par la figure II.1. Cette taxonomie est une extension de celle établie par Guilles et al. [49]. Plusieurs aspects des réseaux ont été investigués dans cette optique, ceci inclus le contenu des messages, l'aspect topologique des graphes sous-jacents des réseaux sociaux, l'aspect temporel des interactions entre les utilisateurs. L'utilisation du contenu des messages

inclus les unités lexicales présentent dans le texte, ou encore des marqueurs tels que les hashtags et les liens hypertextes (URL) qui sont souvent échangés sur les réseaux sociaux. L'aspect sémantique a aussi été exploré, il s'agit d'extraire du texte des messages les thématiques discutées sous forme d'un ensemble d'unité lexicale récurrentes dans un ensemble de tweets.

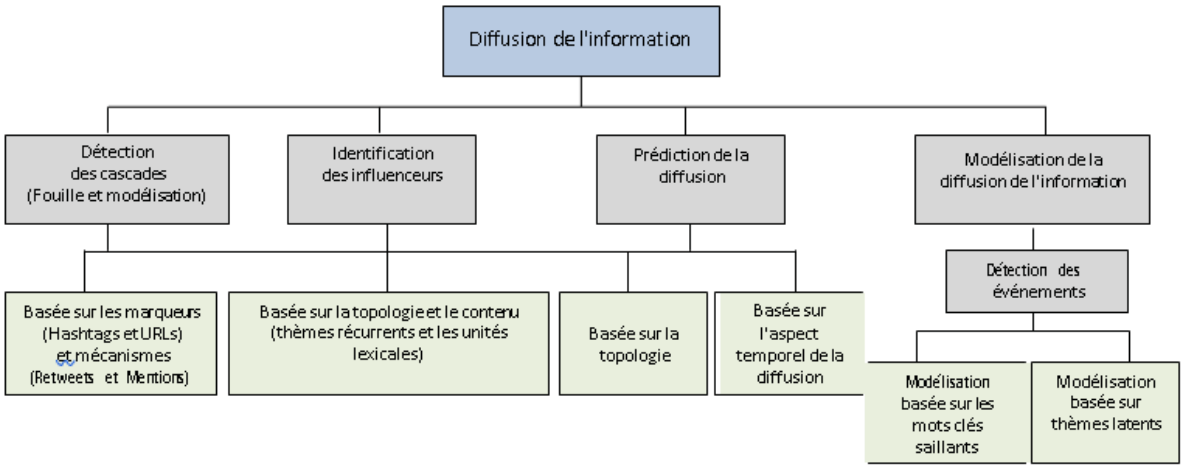

Figure II. 1:Taxonomie des travaux sur la diffusion de l'information

Les solutions proposées pour ces divers problèmes peuvent aussi être catégorisées en deux types, à savoir: les travaux centrés sur l'analyse des données du réseau et ceux basés sur les modèles mathématiques génératifs. En effet, l'analyse permet d'avoir une idée en se basant uniquement sur la fouille des données des réseaux pour en extraire les propriétés de ces derniers, telles que, par exemple, la loi de probabilité qui régie le nombre de suiveurs (followers) des utilisateurs sur Twitter. Par contre, la modélisation permet de fournir des modèles mathématiques qui imitent avec une certaine précision les propriétés des réseaux sociaux.

Bien que beaucoup de travaux ont été entrepris dans ce domaine, un aspect reste encore à explorer à savoir l'utilisation de la sémantique des échanges entre les utilisateurs. En effet, cet aspect est essentielle pour comprendre les propriétés des réseaux sociaux. Par exemple, une application très recherchée au sein de ces derniers est d'identifier les communautés par thématiques. Ceci est d'une grande utilité pour plusieurs acteurs car il permet de cibler un grand nombre d'utilisateurs en même temps en identifiant les profils recherchés. Par ailleurs, l'utilisation de la sémantique permet, entre autres, d'identifier les rumeurs sur les réseaux, la désinformation, etc.

L'extraction de la sémantique des messages reste un challenge vu la nature des messages des réseaux sociaux. Spécifiquement, les messages de Twitter sont courts, ambigus et informel ; ce qui rend l'extraction d'un contexte à partir de ces derniers une tâche difficile. Afin d'atténuer cette difficulté, pour des applications d'analyse des réseaux sociaux, des chercheurs ont utilisé les unités lexicales qui apparaissent dans le texte, c'est à dire, tout le contenu de ces derniers ; ou encore des contenus spécifiques plus facile à appréhender appelés marqueurs, tels que les hashtags et les lien hypertextes. Ces derniers sont considérés comme marqueurs car l'idée consiste à traquer leur cheminement sur le réseau tel un colorant utilisé pour détecter les fuites d'un liquide. Par ailleurs, pour certaines applications d'autres mécanismes ont été explorés spécifiquement sur Twitter, tels que les retweets et les mentions. Ces techniques sont quelque peu efficaces dans la mesure où elles indiquent le crédit accordé aux utilisateurs par le reste des acteurs du réseau. Cependant, leurs utilisation dans le réseau Twitter n'est pas généralisée impliquant ainsi des résultats biaisés.

II.3 Définition du problème

La notion de cascades d'information inclus plusieurs acteurs à savoir: l'élément d'information qui a été propagé, la séquence d'activation; c'est à dire la séquence des utilisateurs qui ont propagé l'information selon un ordre chronologique et sous la condition de la présence d'un lien entre les utilisateurs. Dans ce qui suit, les définitions sont spécifiquement énoncées relativement au réseau social Twitter sur lequel nous portons notre intérêt.

Définition 1. (Séquence d'activation). Une séquence d'utilisateurs U_1, U_2, ..., U_n telle que chaque utilisateur U_i est le suiveur de l'utilisateur U_{i-1} pour i allant de 2 à n.

Définition 2. (Cascade de l'information). Une cascade d'information est une séquence d'activation U_1, U_2, ..., U_n où tous les utilisateurs ont posté le même élément d'information E_j dans l'ordre de la séquence. C'est à dire, si un utilisateur U_i a posté l'élément d'information E_j à l'instant t_k, alors tous les utilisateurs de la séquence situés après Ui auraient aussi posté E_j à l'instant t_l tel que $t_l > t_k$.

Problème (Prédiction des cascades de l'information). Étant donné un élément d'information E_i publié par un utilisateur U_0 à t_0, la prédiction des cascades d'information consiste à prédire la propagation dans le réseau de l'élément d'information E_i, à travers une séquence d'activation $(U_0, U_1, \ldots, U_N)$ à un temps t_p.

Intuitivement, le problème de la prédiction des cascades d'éléments d'information vise à détecter dans un stade précoce les éléments d'information importants, c'est à dire, ceux qui vont être propagés sur le réseau et éventuellement devenir viral. La figure II.2 ci-dessous illustre un exemple de cascade de l'information du message m_8 par le mécanisme de Retweet. Les nœuds et les arcs représentent le graphe sous-jacent du réseau, la pile des messages au niveau de chaque nœud représente la *timeline* de chaque utilisateur et les flèches bleues désignent la cascade d'information du message m_8, et la séquence (U_4, U_5, U_3, U_1) représente la séquence d'activation.

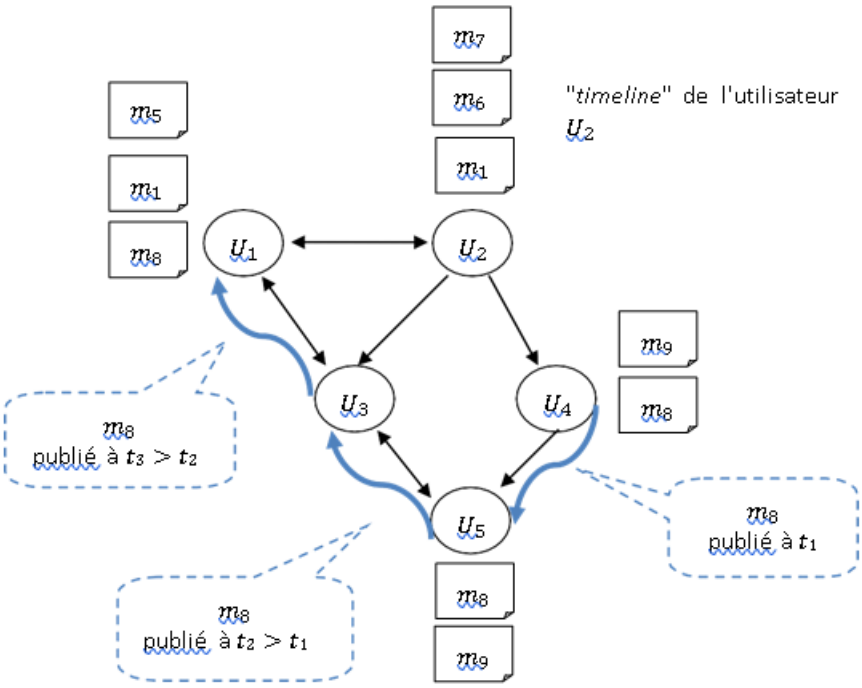

Figure II. 2:Illustration d'une cascade de l'information sur le réseau social Twitter

Le concept de cascades d'information est souvent lié à des problèmes connexes tels que la détection des influenceurs et de la popularité d'un élément d'information. En effet, Karthik et al. [48] ont utilisé les séquences d'activation qu'ils ont appelées chemins fréquents pour la détection des influenceurs. A cet effet, dans leur travaux [48], les auteurs ont proposé un

nouvel algorithme *InFlowMine* pour la détection de toutes les séquences d'activation possibles; pour ensuite les utiliser afin de déterminer les k-top influenceurs. Le problème traité par ces auteurs dans ne concerne pas la prédiction des cascades d'éléments d'information, mais leur détection et leur analyse. Par ailleurs, la définition des séquences d'activation utilisée par les auteurs et quelque peu différentes car les auteurs ont considéré les séquences ayant propagé plusieurs éléments d'information.

II.4 Etat de l'art

Selon [8], trois aspects catégorisent le problème de la prédiction des cascades de l'information. Le premier aspect concerne la nature du problème en lui même, en effet, si ce dernier consiste à prédire la popularité d'un élément d'information; alors le problème est réduit à une régression. Sinon, s'il s'agit à prédire si un élément va être propagé dans le réseau, le problème est une classification. Un autre aspect inhérent au moment de la prédiction proprement, c'est à dire, avant ou après la publication du contenu . Finalement, le troisième aspect concerne la granularité de la prédiction qui peut être effectuée à l'échelle globale (tout le réseau), communautaire (cluster ou communauté) ou individuel (utilisateur).

La prédiction au moment de la publication du contenu est plus difficile à entreprendre, car peu d'information sont disponibles à cet instant. Cependant, cette tâche revêt une plus grande importance dans la mesure où elle offre un avantage pour ceux qui détiennent l'information. Les auteurs [8] propose une taxonomie des travaux de l'art selon cette catégorisation, La figure II.1 ci-dessous illustre une classification des travaux proposé dans [8].

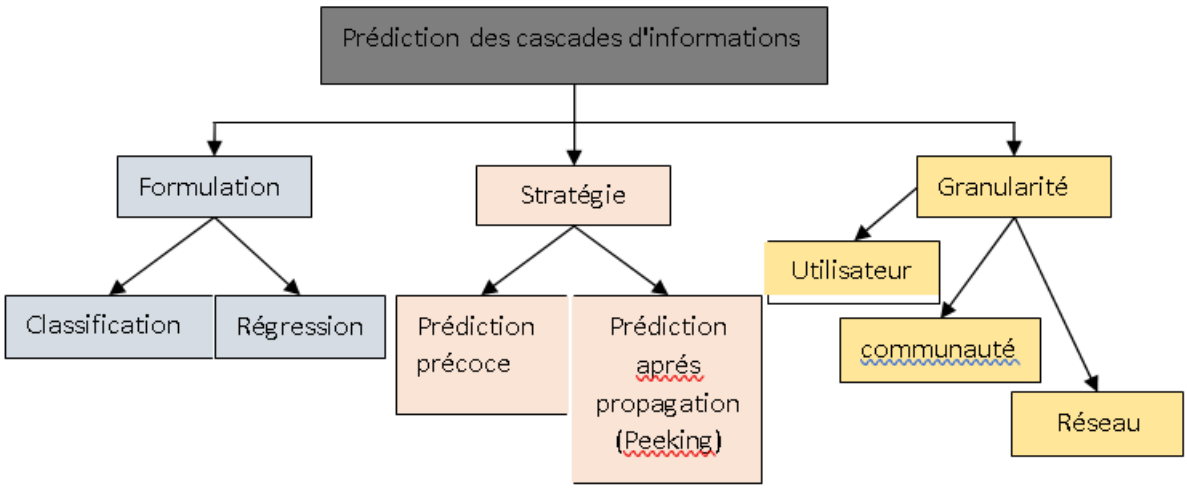

Figure II. 3:Taxonomie des travaux de l'état de l'art selon trois catégorie [8]

Etant donné que le problème étudié peut être mappé à une classification, les algorithmes traditionnels d'apprentissage automatique nécessitant une ingénierie des caractéristiques ont été utilisés, ainsi que l'apprentissage profond. Par ailleurs, des modèles génératifs tels que le modèle de l'épidémie, des survivants et le modèle de poisson ont été explorés. Une taxonomie des travaux, proposée dans [8] , inhérente à la classification est présentée dans la figure II.2.

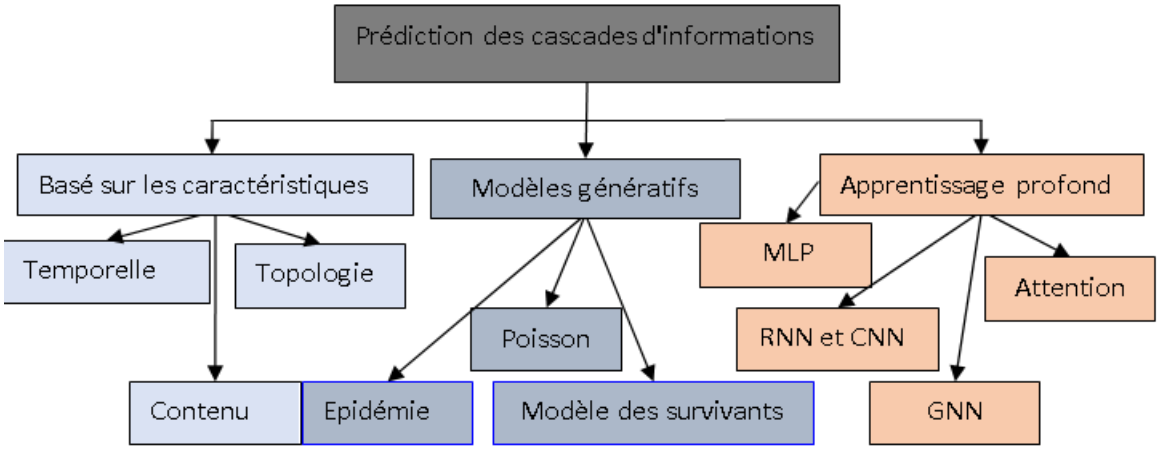

Figure II. 4:Taxonomie des travaux des stratégies de classification étudiées [8]

Plusieurs aspects des réseaux sociaux ont été exploités dans la littérature dont l'aspect temporelle de la publication, la topologie des graphes sous-jacents des réseaux sociaux, et le contenu des messages propagés [8] .

II.4.1 Prédiction basée sur le contenu des messages

Le contenu des publications sur les réseaux sociaux inclut les images, les vidéos et le texte des messages. Certains travaux de la littérature sont focalisés sur les images contenues dans les messages. En effet, parmi les travaux fondamentaux, les auteurs dans [10] se sont intéressés a explorer la corrélation entre certaines caractéristiques des images du réseau social Flickr et leurs nombres de vues. D'autres auteurs ont étudié la prédiction de la diffusion des vidéos. Les auteurs dans [11] ont exploré les caractéristiques artisanales et aussi l'apprentissage profond en utilisant les réseaux de neurones convolutifs (CNN). Dans ce projet, nous nous intéressons entre autres au contenu textuel des messages des réseaux sociaux. En effet, le texte étant le contenu le plus

fréquemment utilisé par les utilisateurs dans certains réseaux, spécifiquement, Twitter qui constitue le réseaux social sur lequel nous nous focalisons dans notre expérimentation.

L'utilisation du contenu dans le domaine de la diffusion de l'information inclue les unités lexicales présentes dans les messages, des unités lexicales spécifiques qui sont utilisées par les utilisateurs de Twitter telles que les hashtags, les liens hypertextes (URLs), ces mécanismes ont été explorés dans [18]. Ces dernières concerne l'aspect syntaxique du texte, en effet, elles sont considérées comme des chaines de caractères et aucun intérêt n'est porté à leurs sens. Ceci est valable aussi pour les hashtags et les URLs, Ces derniers sont communément appelés marqueurs. L'utilisation des hashtags découle du fait qu'ils sont pensés par les utilisateurs pour indexer le texte de leurs messages. D'où, ils sont considérés comme un aspect sémantique du texte. Des travaux plus élaborés utilisent des techniques pour extraire les élément important du texte, telles que la TF-IDF (pour Term Frequency-Inverse Document Frequency) et la LDA (pour Latent Dirichlet Allocation) [8] .

Ces techniques sont utilisés pour extraire les mots clés qui constituent les unités lexicales les plus importantes, ou encore, dans le cas de la LDA, les thématiques des messages sous formes d'un ensemble de mots récurrents. Les auteurs des travaux dans [19,20,21] ont exploré, à travers ces techniques, les distributions des thématiques des tweets pour la prédiction de la popularité des éléments d'information contenus dans les messages. D'autres recherches ont exploité le mécanisme de retweet qui concerne aussi la propagation du contenu dans sa totalité, telles que les travaux dans [18, 22, 23, 24] .

Etant donné que plusieurs réseaux sociaux ont été dédiés aux échanges d'images telle que Pinterest et Instagram, des chercheurs se sont intéressés à utiliser ce support pour la prédiction des cascades. En effet, les travaux dans [25] ont exploité des attributs relatifs aux images tels que l'orientation, la taille, les couleurs, les coordonnées, etc. Dans la même optique d'autres travaux ont explorés la résolution des images, le contraste, la luminosité, etc. [26,27,28,29] . Par ailleurs, le support vidéo a aussi été exploré par une multitude de recherches telles que les travaux dans [29,30,31,32,33] , où plusieurs aspects des vidéos ont été décortiqués pour la prédiction des cascades, tels que la dynamique des scènes et d'autres caractéristiques multimodales.

II.4.2 Prédiction basée sur la topologie

Les graphes sous-jacents des réseaux sociaux donnent beaucoup d'information sur la centralité des utilisateurs (nœuds du graphe). En effet, s'il s'agit du graphe suiveurs/suivis (follower/followee) dans le cas de Twitter, la centralité ou la popularité d'un utilisateur peuvent être estimées par des mesures statiques indiquant à quel point un utilisateur peut fédérer des amis. Un exemple de mesures de centralité est le nombre de triangle du graphe sous-jacent qu'un utilisateur ferme avec ses voisins. La structure du graphe a été exploité dans plusieurs travaux pour la prédiction des cascades de l'information [34,35,36,37,38,39] .Ces travaux ont considérés plusieurs aspects structurels tels que le graphe des cascades en considérant à la fois les utilisateurs ayant pris part aux cascades et également ceux qui n'ont pas participé.

Selon les auteurs dans [8] , les résultats montrent que la corrélation entre la popularité des cascades d'inscription à LinkedIn et la viralité structurelle est élevée, ce qui montre une influence structurel très similaire aux cascades de citations de l'APS. Par ailleurs, les auteurs soulignent le fait que le contenu des éléments dans les messages échangés peuvent affecter la diffusion de l'information.

II.4.3 Prédiction basée sur l'aspect temporel des publication

Les caractéristiques temporelles des éléments d'information et des cascades sont parmi les facteurs les plus importants pour la prédiction de la popularité , nous discutons des caractéristiques temporelles dans le contexte de l'ingénierie des caractéristiques. L'heure du tweet a été utilisée pour éliminer l'effet diurne déséquilibré des activités des utilisateurs, tandis que d'autres facteurs temporels tels que l'heure de Digg, l'heure de la source et la variabilité de l'activité des utilisateurs ont été utilisés pour améliorer la robustesse des modèles. En effet, certains travaux n'explorent que les articles publiés pendant la journée pour entraîner leurs modèles [8] . Par ailleurs, il a été constaté [8] . que le temps de la première participation est un facteur très important pour la prédiction des cascades d'information. En effet, il a été établi que la plupart des tweets ayant reçu au moins 10 retweets en 24 heures ont reçu le premier retweet au plus tard 1 heure après leur publication. Aussi, d'autres facteurs tels que le temps de réaction et la période d'inactivité ont été explorés [8] . Malgré l'importance des caractéristiques temporelles, des études récentes ont également suggéré qu'elles peuvent ne pas fonctionner dans certains cas de figure [15] , en effet, leurs avantages diminuent avec le temps et leur effet n'est pas est moindre comparé aux autres caractéristiques.

II.4.4 Apprentissage automatique et modèles stochastiques

Dans le domaine de la prédiction des cascades plusieurs techniques d'apprentissage ont été explorées afin d'améliorer la précision des systèmes proposés dans la littérature. Nous citerons la régression logistique, la méthode probabiliste naïve de Bayes, les machines à vecteurs supports (SVM) et les arbres de décision. Ces derniers ont été associés à une phase importante d'ingénierie des caractéristiques telles que celles évoquées dans les sections précédentes. D'autres part, les modèles à base de processus stochastiques ont été également utilisés (Poisson et Hawkes). D'autres travaux ont utilisés l'apprentissage profond tels que DeepWalk [40] , node2vec [41] ,les réseaux des graphes convolutionnels (GCN) [42,43] , les réseaux de neurones récurrents (RNN) et leurs variantes tels que les travaux dans [44,45] . Parailleurs les travaux récents, tels que ceux élaborés dans [46,47] s'intéressent beaucoup plus à l'ingénierie des caractéristiques et l'apprentissage automatique classique pour la prédiction descascades.

Chapitre III :
Etude de cas

III.1 Introduction

Dans ce chapitre, nous présenterons une contribution dans la prédiction des cascades de l'information. A cet effet, nous nous sommes intéressés en particulier aux travaux de N. Singh et al. [17] dans l'optique d'améliorer leur approche proposée pour la prédiction des cascades d'information. Dans ce travail de recherche assez récent, les auteurs ont combiné plusieurs caractéristiques de différentes natures dans le but de prédire les cascades d'information dans les réseaux sociaux en ligne.

L'intérêt que nous portons à ce travail réside dans le fait que c'est une tentative de combinaison de l'homophilie et de la similarité du contenu des échanges à la fois. En effet, d'abord, Il a été établie dans la littérature que les utilisateurs qui se ressemblent ont tendance à communiquer [52,53]. Aussi, l'utilisation de la similarité du contenu découle de l'hypothèse que les utilisateurs qui ont les mêmes centres d'intérêts ont plus probablement tendance à socialiser.

Toutefois, nous avons remarqué plusieurs insuffisances auxquelles nous pouvons remédier pour améliorer ce travail.

III.2 Motivation

Les auteurs ont utilisé une approche basée sur la marche aléatoire (Random Walk) et la similarité entre les utilisateurs. Particulièrement, ils ont exploité deux sortes de similarité; la similarité entre les utilisateurs eux-mêmes et la similarité du contenu échangé entre eux. En effet, la similarité entre les utilisateurs qui se traduit par le phénomène d'homophilie gouverne la formation du graphe de relations entre les utilisateurs (voir section I.2.1). Par ailleurs, la similarité du contenu se traduit par l'hypothèse qui stipule si deux utilisateurs ont les mêmes sujets d'intérêt, ils ont tendance à propager les information de l'un et de l'autre. N. Singh et al. ont évalué leur système sur un jeu de test qui consiste en un ensemble de tweets acquis par l'API de Twitter. Cependant, dans leurs travaux, les auteurs ont utilisé un graphe qu'ils ont construit à partir du mécanisme de retweet de Twitter. Ce graphe n'est pas le graphe de relation de Twitter acquis par l'API appropriée (API qui fournit pour un utilisateur donné ces suiveurs), mais construit uniquement à partir des tweets. Qualifié d'arbres de diffusion par les auteurs, chaque nœud du graphe est un utilisateur, l'arc orienté entre chaque deux nœuds signifie que le nœud destination a retweeté le contenu du nœud origine de l'arc orienté.

Le graphe utilisé par N. Singh et al. n'est pas représentatif à notre avis, car d'après une étude sur Twitter [54] il a été établi que seulement 22.4% des tweets sont des retweets. De ce fait, nous estimons qu'il faut utiliser le graphe de relation à travers son acquisition à partir de Twitter par l'API appropriée.

Par ailleurs, la similarité du contenu utilisée dans la méthode de N. Singh et al. [17] consiste à utiliser les unités lexicales contenues dans les messages et à calculer le cosinus de des vecteurs obtenus des échanges entre deux utilisateurs donnés. Toutefois, nous estimons que cette démarche peut être améliorée en ne considérant que les entités importantes du textes d'un message posté. En effet, afin de procéder à une expérimentation pour l'évaluation du système, il indispensable d'avoir une réalité terrain (ground truth); c'est à dire, il faut détecter au préalable toutes les cascades possibles pour procéder à un étiquetage de la base. Cependant, si les bases de tweets sont grandes; il faut que des algorithmes de détection soient scalables sachant que ceux existants focalisent sur les hashtags seulement pour pouvoir contenir les structures de données nécessaires en mémoire. D'aprés une étude sur Twitter [54] il a été établi que seulement 14.5% des tweets contiennent des hashtags, d'où leurs utilisation n'est pas généralisée sur Twitter.

A cet effet, nous préconisons d'utiliser pour l'évaluation de la similarité du contenu les entités nommées comme éléments d'information importants à traquer pour la détection précoce des cascades d'information.

III.3 Cas d'étude

Dans ce projet, nous proposons l'approche proposée par N. Signh et al. [17] ,qui constitue notre cas d'étude. Cette méthode utilise une combinaison de trois caractéristiques : la similarité entre les utilisateurs (homophilie), la similarité du contenu des échanges entre les utilisateurs et la méthode de la marche aléatoire. La marche aléatoire permet de prendre en compte la connexion entre les nœuds. La formulation de chacune des caractéristiques est donnée par ce qui suit:

- **Similitude entre les utilisateurs :** Les auteurs ont adopté les mécanismes de Twitter qui constitue le réseau social sur lequel ils ont conduit leur expérimentation. Si deux utilisateurs ont une relation Follower-Followee ou ont répondu/retweeté à un tweet original, alors ils sont similaires, sinon non.

$$S_{ij} = \begin{cases} 1, & si\ l'utilisateur\ U_i\ est\ similaire\ \text{à}\ U_j \\ 0, & sinon \end{cases} \qquad \text{(III.1)}$$

- **Similitude du contenu :** La similarité du contenu échangé (tweets) est calculée par les auteurs dans [17] en utilisant la similarité en cosinus. A partir de deux tweets un vocabulaire est constitué, puis à chaque tweets t_i est associé un vecteur $\ddot{t}_i$ de longueur égale à la taille du vocabulaire et dont chaque composante contient 1 ou 0 en fonction de l'occurrence d'un item du vocabulaire dans le tweet. Ensuite, la similarité est évalué en utilisant le cosinus entre les deux vecteurs représentant deux tweets de deux utilisateurs donnés. La similitude du contenu est donnée par l'équation:

$$CS_{ij} = \frac{\langle \ddot{t}_i . \ddot{t}_j \rangle}{\|\ddot{t}_i\| \|\ddot{t}_j\|} \quad \text{(III.2)}$$

- **Marche aléatoire (RandomWalk) :** Les auteurs dans [17] ont exploité la marche aléatoire pour prédire les processus de cascade. La marche aléatoire finie est constituée d'un nombre fixe d'étapes. Considérons une marche aléatoire qui part du nœud ni et s'arrête finalement au nœud nj. Si pi représente la probabilité de quitter le nœud i et pi, j la probabilité d'atteindre le nœud j après avoir quitté i. Alors une marche aléatoire partant du nœud i et de longueur n peut être représenté comme :

$$p_i = \{p_{i1}, p_{i2}, \dots, p_{in}\} \quad \text{(III.3)}$$

La probabilité d'être dans un état stationnaire est $p_{ii} = 1 - \alpha$, où α représente la probabilité que le marcheur va quitter avec certitude son état actuel. La probabilité antérieure est proportionnelle au degré sortant du nœud, c'est à dire au nombre de ces suiveurs.

La matrice de similarité et de probabilité est calculée par analogie comme pour les métriques S, CS et P en utilisant les équations III.1, III.2 et III.3. La matrice de degré correspondante est donnée par D. Une fois l'état stationnaire atteint, la probabilité que le marcheur aléatoire reste au nœud j est proportionnelle à la similarité entre deux nœuds et est donnée par :

$$P = DS \quad \text{(III.4)}$$

L'état stationnaire final de la marche aléatoire peut être calculé comme suit :

$$R(t+1) = \alpha PR(t) + (1-\alpha)I \quad \text{(III.5)}$$

$R(t)$, $R(t\ +\ 1)$ sont les matrices de probabilités des états respectifs t et $t + 1$.
Finalement, l'algorithme proposé pat N. Singh et al. [17] pour la prévision de la taille des cascades est donné comme suit:

Algorithme 1.

Entrée : Graphe de relation sous forme de liste d'adjacence

Sortie : Matrice de probabilité des états

1	Générer un graphe représenté par G(V, E) où V représente l'ensemble des nœuds et E l'ensemble des arcs orientés
2	Calculer la matrice de similarité entre les utilisateurs, où chaque cellule S_{ij} représente la similarité entre une de sommets (i, j) en utilisant l'équation III.1;
3	Calculer la matrice de similarité des tweets, où chaque cellule CS_{ij} représente la similarité du contenu pour chaque paire de sommets (i, j) en utilisant l'équation III.2;
4	Calculer la matrice de probabilité de transition P en utilisant l'équation 4;
5	Pour chaque nœud : calculer la probabilité de l'état stationnaire par l'équation 5;

Afin de mieux appréhender la notion de similitude dans ce travail, nous projetons d'utiliser l'extraction des entités nommées pour estimer la similarité du contenu entre les utilisateurs.

III.4 Les entités nommées

L'analyse sémantique des contenus des tweets est une tâche plus difficile que l'analyse du texte riche. En effet, les textes des tweets sont informel, courts et ambigus vu la restriction imposée sur le nombre de caractère des messages par les concepteurs. il devient plus que nécessaire d'exploiter la sémantique des échanges des utilisateurs des réseaux sociaux dont la finalité et de découvrir les thématiques de conversations. Cette tache est généralement basée sur l'exploitation d'un ensemble de tweets et fournit comme résultat un ensemble de termes récurrents pour caractériser une thématique. Cependant, cette démarche considère un thème comme un ensemble de mots récurrents et ne peut extraire le sujet à partir d'un seul tweet. En effet, la définition de la notion de détection des thèmes dans la littérature se traduit par l'identification des

termes récurrents dans une collection de tweets. Par ailleurs, il existe une autre approche qui exploite les données ouvertes et liées ("*Linked Open Data*"). Cette dernière consiste a extraire les entités d'un tweet et de les lier a des ressources d'une base de connaissances, telles que les pages de l'encyclopédie Wikipédia. Cette technique appelée NED ("*Named Entity Disambiguation*") est plus appropriée pour traiter un seul tweet. La détection des entités nommés a évolué depuis l'introduction du concept, de nos jours ce concept est appelé "*Entity Linking*". En effet, la tâche n'est plus restreinte à la détection des entités nommées vu les différentes interprétations du problème qui ont émergées au fil des années. Dans ce qui suit, nous référerons à cette tâche par Extraction et la Liaison d'Entités (ELE), et nous considérerons la définition du problème proposée par M.-C. NAIT-HAMOUD et al. [51] . En effet, la définition proposée par ces auteurs a abouti à la conception d'un système, appelé TELS par les auteurs, pour la détection et l'extraction des entités à partir des tweets plus efficace que les méthodes existantes. La figure III.1 illustre un exemple de ce concept, où il s'agit s'extraire les partie du texte importante dans le texte et de les associer à Wikipédia. Le système TELS prend en entrée un tweet et retourne en sortie les numéro des pages Wikipédia référant aux entités les plus importantes selon l'interprétation proposée par par M.-C. NAIT-HAMOUD et al. dans [51].

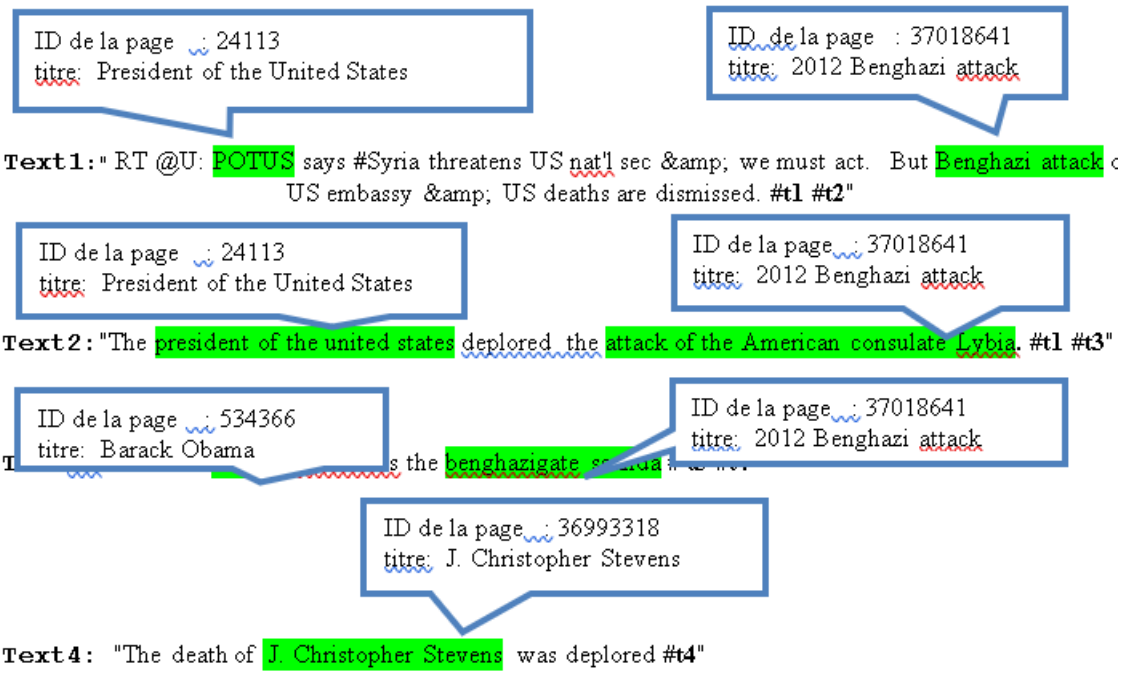

Figure III. 1:Exemple d'extraction des entités à partir de tweets

III.5 Approche proposée

L'approche proposée est basée sur l'apprentissage automatique et une phase d'ingénierie des caractéristiques. La nouveauté de notre travail par rapport aux travaux de N Singh et al. et de considérer les entités extraite par le système TELS [51] comme élément d'information à traquer sur le réseau pour prédire les cascades d'information dans les réseaux sociaux. La figure III.2 illustre notre approche, la base contenant la réalité terrain est décrite dans la section III.6.2 qui concerne l'expérimentation. Le problème de prédiction des cascades des entités est ramené à un problème de classification binaire dans le but et de déterminer si une entité sera propagée ou pas. Dans notre approche, nous avons choisi d'utiliser les forêts aléatoires (Random forest) comme classifieur. Ce choix est justifié par la comparaison effectuée par N. Singh et al.[17] qui établi que ce dernier donne de meilleurs résultats.

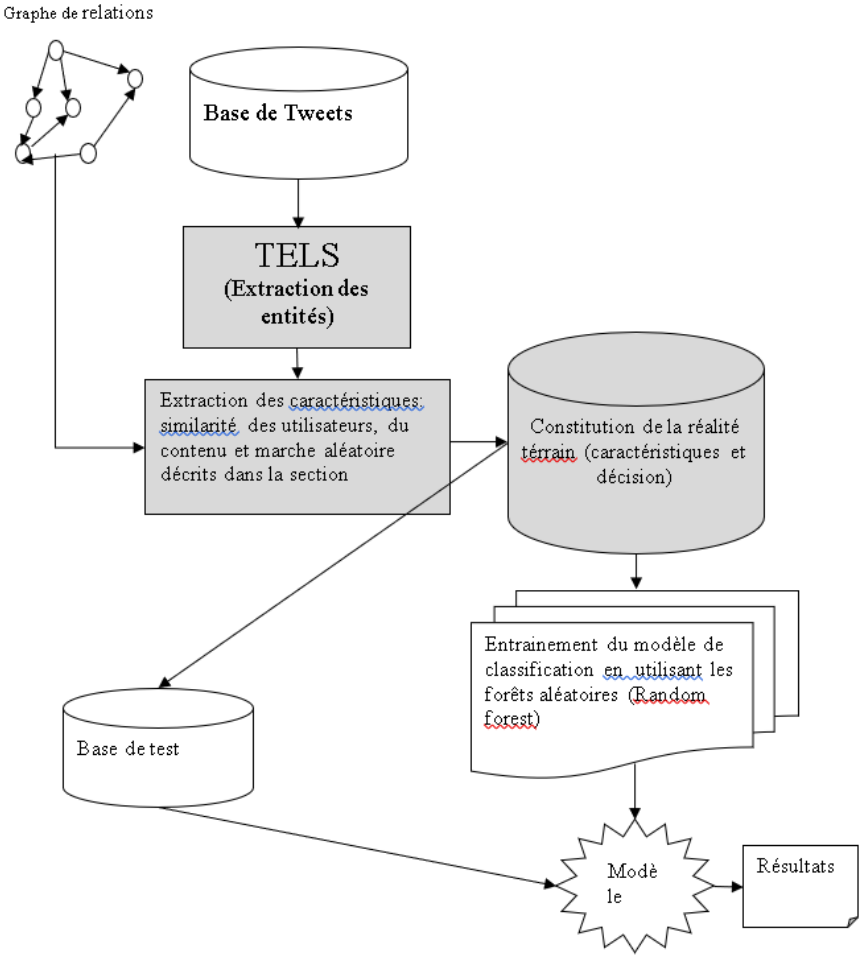

Figure III. 2:Approche proposée

III.6 Conclusion et travaux futurs

Dans ce travail, nous avons expérimenté une approche pour améliorer les travaux des auteurs N. Singh et al. Les insuffisances constatées dans ces derniers concerne beaucoup plus le calcul de la similarité du contenu échangé entre les utilisateurs. A cet effet, nous avons utilisé le concept des entités afin de n'extraire que les informations importantes des tweets, et d'exploiter avec plus l'efficacité la sémantique des messages. Notre deuxième contribution consiste en l'introduction d'une réalité terrain (rendu anonyme) du problème traité. En effet, il n'existe pas de benchmark pour la validation des systèmes à cause des termes de services de Twitter qui interdisent le partage des informations collectées.

Dans nos travaux futurs, nous préconisons de tester les techniques d'apprentissage profond et d'évaluer la nécessité d'une phase d'ingénierie des caractéristiques.

Conclusion générale

Conclusion générale & perspective

Dans cette thèse, nous avons discuté d'une nouvelle méthodologie basée sur la marche aléatoire et les mesures de similarité pour prévoir les cascades sur Twitter. Utilisation des méthodes de traitement du langage naturel sur un ensemble de données réelles recueillies à partir de Twitter.

Nos résultats montrent que le modèle fonctionne mieux.

Pour les besoins de ce travail, nous avons utilisé le modèle de marche aléatoire de base. Nous voudrions exécuter nos tests en utilisant des ensembles de données plus significatifs dans les recherches futures. De plus, nous aimerions incorporer d'autres méthodes d'ensemble, telles que des algorithmes de boosting, pour améliorer la prédiction en cascade. Nous ajouterons également plus de mesures de similarité au modèle pour l'enrichir.

Références Bibliographies

Références Bibliographies

[1] https://sysomos.com/inside-twitter/twitter-friendship-data

[2] The strength of weak ties in business – Research Hub (alexanderchekanov.com)

[3] De Choudhury, M. (2011, October). Tie formation on twitter: Homophily and structure of egocentric networks. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing* (pp. 465-470). IEEE.

[4] Yamaguchi, K., & Kandel, D. (1993). Marital homophily on illicit drug use among young adults: Assortative mating or marital influence?. *Social Forces*, *72*(2), 505-528.

[5] Planetoscope - Statistiques : Nombre de tweets exp**di**s sur Twitter

[6] Brandwatch : 114 statistiques et faits impressionnants à retenir sur les médias sociaux | Brandwatch

[7] Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press.

[8] Zhou, F., Xu, X., Trajcevski, G., & Zhang, K. (2021). A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)*, *54*(2), 1-36.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. JMLR 3(Jan. 2003), 993–1022

[10] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. 2011. Predicting popular messages in twitter. In WWW Companion

[11] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular? In WWW. 867–876

[12] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In MM. 898–907

[13] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and authorcontrolled natural experiments on twitter

[14] Yu Yang and Jian Pei. 2019. Influence analysis in evolving networks: A survey.

TKDE (2019), 1–19.

[15] Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In WSDM. 177–186.

[16] Daniel Xie, Jiejun Xu, and Tsai-Ching Lu. 2017. What's trending tomorrow, today: Using early adopters to discover popular posts on Tumblr. In IEEE BigData. 2168–2176

[17] N. Singh, A. Singh, and R. Sharma. "Predicting Information Cascade on Twitter Using Random Walk". In: Procedia Computer Science. Vol. 173. 2020. DOI: 10.1016/j.procs.2020.06.024.

[18] Fan Zhou, Xovee Xu, Kunpeng Zhang, Goce Trajcevski, and Ting Zhong. 2020. Variational information diffusion for probabilistic cascades prediction. In INFOCOM. 1618–1627.

[19] Su-Do Kim, Sung-Hwan Kim, and Hwan-Gue Cho. 2011. Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity. In CIT. 449–454.

[20] Tae Yano and Noah A. Smith. 2010. What's worthy of comment? Content and comment volume in political blogs. In ICWSM

[21] Tauhid R. Zaman, Emily B. Fox, Eric T. Bradlow, et al. 2014. A bayesian approach for predicting the popularity of tweets. Ann. Appl. Stat. 8, 3 (2014), 1583–1611.

[22] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten De Rijke. 2012. Predicting imdb movie ratings using social media. In ECIR. 503–507

[23] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In WWW. 591–600.

[24] Karthik Subbian, B. Aditya Prakash, and Lada Adamic. 2017. Detecting large reshare cascades in social networks. In WWW. 597–606.

[25] Matúš Medo, Manuel S. Mariani, An Zeng, and Yi-Cheng Zhang. 2016. Identification and impact of discoverers in online social systems. Sci. Rep. 6 (2016), 34218

[26] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In ICLR.

[27] Andrey Kupavskii, Alexey Umnov, Gleb Gusev, and Pavel Serdyukov. 2013.

Predicting the audience size of a tweet. In ICWSM

[28] Eric Gilbert. 2013. Widespread underprovision on reddit. In CSCW. 803–808.

[29] Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. 2009. Predicting the volume of comments on online news stories. In CIKM. 1765–1768.

[30] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In MM. 898–907

[31] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In MM. 223–232.

[32] Minh X. Hoang, Xuan-Hong Dang, Xiang Wu, Zhenyu Yan, and Ambuj K. Singh. 2017. GPOP: Scalable group-level popularity prediction for online content in social networks. In WWW. 725–733.

[33] Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation count prediction: Learning to estimate future citations for literature. In CIKM. 1247–1252

[34] Peng Bao, Hua-Wei Shen, Junming Huang, and Xue-Qi Cheng. 2013. Popularity prediction in m

[35] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted? In WWW. 925–936.icroblogging network: A case study on sina weibo. In WWW. 177–178.

[36] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. 2010. Outtweeting the twitterers-predicting information cascades in microblogs. Workshop Online Soc. Netw. 10 (2010), 3–11

[37] Shuai Gao, Jun Ma, and Zhumin Chen. 2014. Effective and effortless features for popularity prediction in microblogging network. In WWW. 269–270.

[38] Ali Zarezade, Ali Khodadadi, Mehrdad Farajtabar, Hamid R. Rabiee, and Hongyuan Zha. 2017. Correlated cascades: Compete or cooperate. In AAAI. 238–244

[39] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided hierarchical attention network for multimodal social image popularity prediction. In WWW. 1277–1286

[40] Henrique Pinto, Jussara M. Almeida, and Marcos A. Gonçalves. 2013. Using early view patterns to predict the popularity of youtube videos. In WSDM. 365–374

[41] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. Zighed. 2013. Information diffusion in online social networks: A survey. ACM Sigmod Record 42, 2 (2013),

17–28.

[42] Ryota Kobayashi and Renaud Lambiotte. 2016. TiDeH: Time-dependent hawkes process for predicting retweet dynamics. In ICWSM. 191–200.

[43] Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In WSDM. 177–186.

[44] Junyoung Chung, Caglar Gulcehre, Kyung Hyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. Retrieved from https://ar

[45] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. 2011. Predicting popular messages in twitter. In WWW Companion.

[46] Ruocheng Guo and Paulo Shakarian. 2016. A comparison of methods for cascade prediction. In ASONAM. 591–598.

[47] Luam Catao Totti, Felipe Almeida Costa, Sandra Avila, Eduardo Valle, Wagner Meira Jr., and Virgilio Almeida. 2014. The impact of visual attributes on online image diffusion. In WebSci. 42–51

[48] Karthik Subbian, Charu Aggarwal, and Jaideep Srivastava. "Mining influencers using information flows in social streams". In: ACM Transactions on Knowledge Discovery from Data 10.3 (2016). ISSN: 1556472X. DOI: 10.1145/2815625.

[49] Adrien Guille. "Information diffusion in online social networks". In: Proceedings of the ACM SIGMOD International Conference on Management of Data. 2013. DOI: 10.1145/2483574.2483575.

[50] Wojciech Galuba et al. "Outtweeting the Twitterers - Predicting Information Cascades in Microblogs". In: Proceedings of the 3rd Wonference on Online Social Networks. WOSN'10. USA: USENIX Association, 2010, p. 3.

[51] Mohamed Cherif Nait-Hamoud, Fedoua Lahfa, and Abdellatif Ennaji. "A step further towards a consensus on linking tweets to Wikipedia". In: Evolutionary Intelligence (2021). ISSN: 18645917. DOI: 10.1007/s12065-020-00549-8.

[52] Alan Mislove et al. "You are who you know". In: 2010. DOI: 10.1145/1718487. 1718519.

[53] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. "Birds of a feather: Homophily in social networks". In: Annual Review of Sociology 27 (2001). ISSN: 03600572. DOI: 10.1146/annurev.soc.27.1.415

[54] Hai Liang and KingWa Fu. "Testing propositions derived from twitter studies: Generalization and replication in computational social science". In: PLoS ONE 10.8 (2015). ISSN: 19326203. DOI: 10.1371/journal.pone.0134270.