République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Electronique et d'Informatique
Département Informatique

Mémoire de projet de fin d'études de Master

Option

Ingénierie des Logiciels

Thème

Approche basée sur le contenu pour la recommandation d'influenceurs dans les réseaux sociaux : Application pour les compagnes de marketing

---

Sujet proposé par :

Mr K.BOUKHALFA

Réalisé par :

Melle BELHANAFI Kenza

Melle KHENNAK Sara

Devant le jury composé de :

Mr DAOUDI Président

Mr DJOUADI Membre

Binôme N° : 08/2016

Résumé

Avec le développement des médias sociaux et l'accroissement sans cesse d'opinion disponible en ligne, en particulier sur les réseaux sociaux, il devient de plus en plus utile d'aider les entreprises à suivre les influenceurs qui diffusent des contenus capables de modifier le comportement d'une communauté de lecteurs ou de "suiveurs".

L'objectif de notre travail est la détection des influenceurs sur le réseau social Twitter, à travers l'analyse du contenu de leurs publications. Au début, nous avons procédé à l'extraction des données du réseau social à l'aide de l'API du réseau, et cela par le biais de mots-clés représentant des marques d'un domaine.

Une fois les publications extraites, une phase d'analyse de sentiment du contenu aura lieu, en utilisant une classification supervisée, afin de déterminer si le post dégage une opinion positive ou négative. A la fin, nous procédons à l'extraction des utilisateurs posteurs des tweets pour désigner la source de critique, suivant deux critères d'influence : la popularité et l'impact de l'utilisateur dans le réseau social.

Notre outil de détection d'influenceurs sur Twitter a été développé en utilisant le langage Python, le langage R, ainsi qu'une base de données NoSQL orientée graphe : Neo4j.

Mots clés : Analyse de sentiment, influenceur, réseaux sociaux, classification.

Abstract

With the development of social media and increased constantly available opinion online, especially on social networks,it becomes veryuseful to help companies track the influencers who broadcast content that can change the behavior of readers or "followers" inside a community.
The aim of our work is to detect influencers on the social network Twitter, through the analysis of the content of their publications. At first we made the extraction of social network data using the Network API, and this through keywords representing brands of a domain.

Once publications extracted, the analysis phase of a content feeling will take place, using a supervised classification to determine if the post releasesa positive or negative opinion. At the end, we proceed to the extraction of posters tweets to find out the source of criticism, according to two criteria of influence: the popularity and impact of the user in the social network.

Our influencers detection tool on Twitter was developed using the Python language, the R language and a NoSQL database oriented graph: Neo4j.

Keywords: sentiment analysis, influencer, social networks, classification.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Contexte

L'entreprise est un acteur économique qui permet de produire des biens et des services aptes à satisfaire les besoins des clients, et qui doivent faire face à une concurrence accrue avec sa capacité à maintenir ou à accroître ses parts de marché.
Néanmoins, suite à la vaste offre sur le marché, le consommateur est devenu plus mature et plus exigeant. La segmentation du marché, méthode traditionnelle des entreprises, ne suffit plus à satisfaire le client, qui exige aujourd'hui une offre personnalisée.
De ce fait, avec la naissance d'internet, les réseaux sociaux sont devenus un outil incontournable permettant de renforcer la visibilité des entreprises dans les campagnes marketing, de gérer leur image, de se développer financièrement, de favoriser la gestion de leur e-réputation et d'enrichir leur expérience sur le marché du web qui offre de nouvelles opportunités, auparavant inexistantes. Ce qui fait que dans la démarche marketing du futur, les médias sociaux deviennent un puissant élément créateur d'offres, provoquant des succès commerciaux exceptionnels avec des coûts réduits. Mais également peuvent présenter plusieurs risques d'échecs qui peuvent nuire à l'image de l'entreprise.

Problématique

De nos jours, Internet est devenu un média où l'information se propage très rapidement. Les gens le considèrent comme le média d'information le plus utile, d'où les entreprises ont rapidement tiré profit de cette outil afin de diffuser leurs offres et de développer leurs images de marque, ainsi augmenter leurs parts du marché et leurs réputations au sein du grand public.
L'apparition des médias sociaux a considérablement changé le visage d'internet en instaurant un nouveau paradigme de la communication et de l'échange d'opinion en temps réel, ce qui permet aux internautes d'être connectés en permanence pour analyser les tendances, ceci donne aux entreprises l'habileté d'être ultra réactive.
La montée en puissance des réseaux sociaux sur internet a bousculé les modèles traditionnels de communication des entreprises. Plusieurs millions de personnes sont dorénavant interconnectées et peuvent échanger des discussions sur une infinité de sujets autour des marques qui peuvent être source de notoriété fortuite pour l'entreprise, ou, au contraire décrédibiliser son image de marque et sa réputation pour longtemps. Or les entreprises ne sont pas exemptes de ces échanges.
Nombreuses sont les entreprises victimes de crise de communication online, affectant de manière durable leur image de marque. L'exemple le plus probant est la chute vécu par **Orascom Telecom** « Djezzy » en 2009 suite au match Égypte- Algérie qui a engendré le boycott de la marque sachant que tout à commencer dans les réseaux sociaux.
En outre, les réseaux sociaux tels que Facebook et Twitter ont été d'une très grande responsabilité dans les révolutions arabes au Moyen-Orient. [ref, h]
Le vrai challenge pour les entreprises d'aujourd'hui est de comprendre les spécificités de ces nouveaux médias et d'établir une stratégie de communication efficace, car ses derniers constituent des

plateformes où les internautes évoluent dans leurs sphères d'influence qui représente leurs cercles d'amis. Il est vrai qu'on est plus facilement influençable par nos amis que par les entreprises, car un individu qui évolue dans une société fait ses choix en partie grâce à son entourage notamment avec un système de bouche à oreille, comme l'explique aussi bien **Jeff Bezos**, le PDG de la boutique en ligne **Amazon** « Si vous rendez vos clients mécontents dans le monde réel, ils sont susceptibles d'en parler chacun à 6 amis. Sur internet, vos clients mécontents peuvent en parler chacun à 6000 amis ».[ref, c]
Ainsi, il est important de détecter et suivre les influenceurs, car l'influenceur est un individu qui par son statut et sa capacité de persuader et inciter d'autres individus à adopter certains comportements peut les influencer à changer les comportements de consommation dans un univers donné.
Afin de remédier au problème d'influenceurs, plusieurs techniques ont été proposées pour faciliter la recherche des personnes influentes, en se basant sur les statistiques à travers les relations de rediffusion des messages en tenant compte de la popularité de l'utilisateur. Or, ces travaux permettent d'identifier des influenceurs de façon fortuite ne prenant pas en compte si réellement cet individu peut être classé comme un potentiel influenceurs vu le contenu de sa publication.
D'un autre côté, plusieurs travaux ont été proposés sur la diffusion de l'information et la fouille d'opinion dans les réseaux sociaux basés soit sur le lexique, la syntaxe ou la sémantique de la publication. Néanmoins, toutes ses approches font abstraction de l'influence qui peut porter ces publications sur les clients et de la source de problème, alors que les entreprises sont intéressées de connaitre et de détecter les critiqueurs sources.
De ce fait, l'objectif de notre travail est de proposer une approche qui aide à détecter les influenceurs dans les réseaux sociaux, en se basant sur une analyse de sentiment des publications.

Objectif du mémoire

L'objectif de notre travail est de proposer une approche pour la détection d'influenceurs dans les réseaux sociaux. Nous procédons d'abord à la recherche et la récupération des publications selon certains mots clés. Puis on se base sur le contenu des publications afin d'analyser le sentiment pour but de voir si c'est un poste positif ou négatif, ensuite on sélectionne les influenceurs suivant des statistiques, tel que la popularité de l'utilisateur et son impact dans le réseau social.

Organisation du mémoire

Le présent mémoire est organisé en deux parties distinctes :

- La première partie représente l'état de l'art, qui est organisé en deux chapitres :
  - Le chapitre 1 est consacré à la détection d'influenceurs dans le web sociale pour le marketing viral.
  - Le chapitre 2 présente l'importance de l'analyse de sentiment pour l'identification des influenceurs dans les réseaux sociaux.

- La deuxième partie de notre mémoire, s'articule sur trois chapitres :
  - Le premier chapitre est consacré à la conception de l'approche, que nous proposons pour la détection d'influenceurs dans les réseaux sociaux basé sur le contenu.
  - Le second chapitre présente la conception de notre outil, qui sera basé sur l'UML.
  - Le troisième chapitre présente notre implémentation, où nous présenterons notre outil, en détaillant les choix techniques pour sa réalisation, pour but de valider notre approche.
    Pour finir ; nous clôturons notre mémoire par une conclusion générale de notre travail suivi des perspectives.

Première partie

Etat de l’art

CHAPITRE I

Détection d'influenceurs dans le web social pour le marketing viral

Internet est devenu un outil d'information et de communication nécessaire, permettant d'échanger, de travailler, de rencontrer, d'apprendre et même de commercer, surtout avec la naissance des médias sociaux, où chaque individu devient un média en soi, permettant ainsi de faire valoir ses idées à un grand nombre de personnes.
Néanmoins, le web social devient incontournable dans la stratégie commerciale pour le marketing de toute entreprise, qui permet de développer l'image de marque, d'augmenter leurs ventes et de diffuser des informations et des produits à une échelle mondiale. De plus, les gens doués avec les médias sociaux sont capable de devenir des leaders d'opinions et peuvent avoir une grande influence sur son entourage.
De ce fait ; il est important de détecter ces influenceurs afin de fournir des opportunités pour les campagnes de marketing des entreprises.
Dans ce premier chapitre, nous allons définir le marketing viral, et les raisons qui poussent les entreprises à utiliser ce genre de stratégie pour promouvoir leurs produits. Par la suite nous parlerons des médias sociaux, et plus précisément des réseaux sociaux. A la fin nous présenterons l'influence, ainsi que la détection des influenceurs et leur impact sur l'image de l'entreprise.

I.1 Marketing Viral dans l'entreprise

De tout temps, le consommateur est surexposé à des publicités fournis par les campagnes de marketing des différentes entreprises, qui fait du marketing l'outil indispensable pour se démarquer de la concurrence. Selon [ref, b] Le marketing désigne l'ensemble des méthodes et des actions coordonnées, qui cherchent à déterminer l'offre de produits et de services d'une entreprise, et concourent à leur développement en fonction des attentes et attitudes des consommateurs, et à en faciliter la commercialisation dans les meilleures conditions de profit basé sur la connaissance du marché.
Avec l'introduction d'internet, devenu le premier média en terme de décision d'achat dépassant les médias traditionnels tel que, la télévision et la radio. Le marketing viral est apparu et, il est devenu un sujet très débattu dans l'environnement social vu son importance, créant ainsi de la valeur perçue par les clients et adapte l'offre commerciale de l'entreprise aux désirs des consommateurs.

Le marketing viral ou buzz marketing repose sur un effet de contamination de la cible par la propagation du message selon des techniques et des vecteurs recourant essentiellement au bouche-à-oreille. Tout comme un virus qui se fait discret, Ensuite, lorsque le milieu est favorable, il prend de l'ampleur et plus rien ne peut l'arrêter.
De plus, [E and K.B, 2002] expliquent que : « L'élément stratégique [...] est d'attirer l'attention du consommateur et d'en faire en même temps un agent de communication, autrement dit, permettre à un récepteur de devenir émetteur. Pour cela, il faut que l'intérêt du consommateur s'identifie à l'intérêt du service ou de la marque. C'est cette implication des consommateurs qui permet de réaliser une campagne exponentielle rapide. »
Néanmoins, le marketing viral apparait comme un enjeu stratégique de taille pour les entreprises. Il leurs permet de gagner en termes de coûts ; car l'audience initiale est obtenue « gratuitement », vu que la diffusion est principalement faite sur Internet, ainsi, de tisser des relations particulières avec le consommateur et de susciter chez lui l'intérêt à l'égard de la marque. Cependant, cette technique de communication doit être utilisée avec précaution vue qu'elle peut entrainer certains risques pour l'entreprise, tel que, le détournement du message publicitaire : Les internautes qui souhaitent s'amuser ou causer des dommages à un produit et/ou à une entreprise peuvent détourner ou modifier le message d'une campagne publicitaire. Cette déformation peut aussi être involontaire mais aboutir à des résultats tout autant désastreux [J, 2011]. Et cela, surtout que les médias sociaux sont devenus une partie nécessaire dans la vie de l'entreprise.
Cette stratégie rendue possible grâce à internet est devenue de plus en plus intéressante avec l'apparition des médias sociaux, qui facilitent considérablement ce genre de marketing et aide l'entreprise à la prise de décision.

I.2 Les médias sociaux

Le réseau Internet, par ses caractéristiques et ses capacités technologiques, est un outil de communication très performant. Et comme le développement actuel de son utilisation l'a rendu incontournable, toutes les organisations, quelles que soient leurs champs d'activité, doivent intégrer aujourd'hui ce média dans leur stratégie de communication.
L'apparition des médias sociaux a, en revanche, donné une nouvelle vision d'Internet qui revient à considérer l'internaute comme acteur, et partie prenante de l'information. Le terme « médias sociaux » a été défini par [A.M and M, 2010] comme étant « un groupe d'applications en ligne qui se fondent sur l'idéologie et la technique du web 2.0 et permettent la création et l'échange du contenu généré par les utilisateurs ».
En effet, les médias sociaux se voient comme un service en ligne qui rassemble un ensemble d'outils, et grâce à sa popularité les internautes l'adopte pour partager l'information, et donnent également la liberté à d'autres utilisateurs de la rediffuser et de réagir sur cette même information. Parmi ses outils, nous citons : les blogs, les forums, les wikis, les réseaux sociaux ... etc.

I.2.1 Les Réseaux sociaux

Les réseaux sociaux sont un ensemble des sites internet, permettant de constituer des regroupements virtuels de personnes physiques ou de personnes morales (associations, entreprises, institutions) et fournissent à leurs membres des outils et des interfaces d'interactions, de présentation et de communication, qui leurs permettent de discuter, d'échanger ou bien générer du business, faire connaître leurs entreprises, leurs services ... etc.
Le concept des réseaux sociaux était de valoriser les liens et les relations que peuvent avoir les entités sociales (individus) entre eux, sans tenir compte du rôle qu'occupe chaque individu dans la société. [Donati, 1994]

Selon une étude publiée sur le Journal "Computer-Mediated Communication", les réseaux sociaux sont définis comme étant « l'ensemble des services du Web qui permettent aux individus de construire un profil public ou semi-public dans une communauté virtuelle, articulé d'une liste d'autres utilisateurs avec lesquels ils partagent une connexion, la nature de cette connexion varie d'une communauté à une autre. (Relation d'amitié sur Facebook, relation d'abonnements/abonnés sur Twitter,..) » [N.B,]. Le point fort des réseaux sociaux, c'est leur immédiateté et leur rapidité, car ils constituent un fabuleux accélérateur pour la diffusion d'information. Il suffit qu'un internaute découvre un produit, un service ou une personne qui l'intéresse pour qu'il en informe en temps réel les autres membres de son réseau.

Les réseaux sociaux se sont installés petit à petit dans notre quotidien, bouleversant les méthodes de communication et le partage d'informations. Ces plateformes sociales ne cessent d'évoluer, offrant toujours plus de nouvelles fonctionnalités et basculant de simples plateformes d'interactions à des outils business indispensables.

I.2.1.1 Évolution des réseaux sociaux

Nous allons parler du phénomène mondial qui est né aux États-Unis peu de temps après l'explosion d'Internet dans les années 1990.
Tout a commencé en 1994 lorsque Justin Hall décide de lancer son site "Justin's Links from the Underground" ,pour se connecter au monde extérieur. Hall a publié sur son blog pendant 11 ans et est reconnu comme le père fondateur des blogs personnels.
Le phénomène se propage, et donne naissance en 1995 au réseau Classmates, qui permettait aux américains de retrouver leurs anciens camarades d'école et d'université, l'ancêtre de Copains d'Avant en quelque sorte, néanmoins celui-ci n'offre pas toutes les fonctionnalités des réseaux sociaux actuels.
Ce problème a tout de suite été résolu en 1997, en donnant jour au premier réseau social Sixdegrees, le premier a réuni toutes les fonctionnalités de base d'un réseau sociale, qui permet aux utilisateurs la création de profils et la gestion des listes d'amis. Mais malgré les millions d'utilisateurs, le service a échoué en 2001.

Après l'année 2000 une nouvelle vague est apparue tournée vers le développement de réseaux d'affaires avec le lancement de Ryze en 2001.
A partir de 2003, la création de nombreux sites de réseaux sociaux, a donné lieu à l'emploi du terme YASNS : « YetAnother Social Networking Service » « encore un autre réseau social ».
Parmi ces nouveaux réseaux nous citons trois grands sites, qui font leur apparition entre 2002 et 2003 et qui ont révolutionné notre façon d'utiliser le web, que ce soit dans la sphère privée ou dans la sphère professionnelle : Friendster, WordPress, LinkedIn et bien sur MySpace, qui a réuni 1 million d'utilisateurs en seulement un mois.

Mais ce n'est qu'en 2004, que le réseau le plus populaire est lancé par Mark Zuckerberg. Tout d'abord Facebook apparait comme un service fermé réservé aux membres de l'université de Harvard seulement, ensuite peu à peu a été amélioré et aujourd'hui est devenu un des premiers sites de réseau social généraliste utilisé dans le monde. En 2015, il comptait environ 1.49 milliard d'utilisateurs actifs par mois.
L'utilisation des réseaux sociaux est devenue, de plus en plus intéressante, et leur création ne cesse de se développer et de façon rapide.
Actuellement, les réseaux sociaux sont devenus des outils de communication incontournables pour les entreprises à tous les niveaux, de la promotion de nouveaux produits à la recherche de nouveaux consommateurs, en passant par la fidélisation de leur clientèle,avec un moindre coût et un retour sur investissement beaucoup plus intéressant.

I.2.1.2 Importance des réseaux sociaux dans l'entreprise

La présence des entreprises et des marques sur les réseaux sociaux est de plus en plus généralisée, car ils ont prit conscience de son importance, et ils ont décidé de les utiliser à des fins professionnelles. D'une part, pour gérer leurs e-réputation et d'autre part être plus proche de la clientèle. Ainsi, les entreprises diffusent des publications, tel que des articles, des nouveaux produits, ou autres dans le but de communiquer à ces cibles les dernières initiatives menées par la marque. Et, en terme de promotion, d'autres informations viennent à stimuler les ventes : réductions, ventes privées ... etc leurs succès est garanti par une diffusion efficace et leurs ventes supplémentaires sont susceptibles d'avoir lieu en ligne ou dans un point de vente traditionnel.
D'une manière générale, on peut dire que les réseaux sociaux, qui sont directement rattachés au marketing et à la communication, ont une influence sur absolument tous les types de vente des entreprises. Ces réseaux sont une véritable vitrine virtuelle pour les biens et services.

I.2.2 Les médias sociaux et la technologie NoSQL

Ces dernières années ont été marquées par l'explosion des médias sociaux sur Internet, augmentant le nombre d'usagers sur les différentes plateformes, induisant un nouveau mode de communication et d'interaction sociale.
Ces usagers se retrouvent à la fois auditeurs et locuteurs, échangeant des avis ou des expériences

sur des sujets divers, provoquant un flux d'information bidirectionnel.
Cette évolution des usagers et l'interaction entre les internautes entraînent une augmentation massive du volume de données sur ces plateformes, provoquant une difficulté de gestion, ce qui rend nécessaire l'utilisation de nouveaux moyens technologiques, pour faire persister ce grand volume de données, et pouvoir les interroger et les traiter. Pour cela, les bases de données NoSQL sont une solution à ce problème.

I.2.2.1 Définition du NoSQL

Le **NoSQL** l'acronyme de « Not Only SQL » est une catégorie de bases de données apparues en 2009, qui se différencient du modèle relationnel que l'on trouve dans les bases de données traditionnelles, car elle englobe une étendue de technologie qui lui permet de résoudre des problèmes de performance en matière d'évolutivité et de Big-data.
Cette catégorie de produits fait le compromis d'abandonner certaines fonctionnalités classiques des SGBDs relationnels au profit de la simplicité, la performance et une forte scalabilité. La scalabilité est la capacité d'un système à répondre à une demande toujours croissante de la part des utilisateurs en termes de requêtes [L, 2012].
Aujourd'hui le terme NoSQL englobe tous les SGBDs qui ne suivent pas la tendance de type relationnel. Cela signifie que le NoSQL n'est pas un seul produit ou même une technologie unique. En effet, il représente de nombreux produits ainsi que plusieurs concepts de stockage et de manipulation de données.

I.2.2.2 Types de base de données

Il existe différents types de bases de données NoSQL spécifiques à différents besoins, mais parmi la large gamme on distingue ces 4 grandes familles que nous citons [E, 2012] :

- **Les bases de données clé-valeur :** La plus simple des bases de données NoSQL, cette structure est très adaptée à la gestion des caches, ou pour fournir un accès rapide aux informations. Les données sont donc simplement représentées par un couple clé-valeur. La valeur peut être une simple chaine de caractères ou un objet sérialisé.
- **Les bases de données orientées documents :** La plus adaptée des bases de données NoSQL pour le web, sa représentation est très proche de la représentation clé-valeur, sauf que la valeur dans ce cas, est un document de type JSON ou XML par exemple. L'avantage est de pouvoir récupérer, via une seule clé, un ensemble d'informations structurées de manière hiérarchique.
- **Les bases de données orientées colonnes :** La plus proche des bases de données NoSQL à la présentation d'une table dans un SGBDR, avec une différence est que dans une BD NoSQL orientée colonne, le nombre de colonnes est dynamique, ce qui la rend beaucoup plus évolutive et flexible.

• **Les bases de données orientées graphes :** La plus adaptée des bases de données NoSQL au traitement des données des réseaux sociaux, elle se base sur la théorie des graphes, en s'appuyant sur la notion de nœuds, de relations et de propriétés qui leur sont rattachées.
Ce modèle permet de stocker et de manipuler des données complexes liées par des relations non-triviales ou variables, cependant la base de données orienté graphe la plus connu est Neo4j.

I.2.3 Statistiques actuelles sur les médias sociaux

Les médias sociaux réussissent à rassembler un très important nombre d'internautes, ils leurs permettent d'échanger des avis, des opinions, mais aussi de partager et communiquer avec leur entourage, ainsi réduisant la distance et renforçant la sociabilité des individus.
Cependant la facilité et l'accessibilité de ces médias ont déclenché une suite de changement dans les habitudes des internautes, qui sont devenus accro à ses derniers.
Les statistiques actuelles parlent eux-mêmes de l'importance qu'occupent aujourd'hui les réseaux sociaux. Selon le blog de modérateur, les chiffres clés des réseaux sociaux en 2016 ainsi [ref, d] :

- Sur 7,357 milliards de personnes dans le monde, on dénombre 3,715 milliards d'internautes.
- Sur 3,715 milliards d'internautes, 2,206 milliards utilisent les réseaux sociaux chaque mois.
- Sur 3,715 milliards d'internautes, 68% des internautes et 28% de la population mondiale, consomment 2h par jour comme temps moyen de connexion pour les réseaux sociaux.

Selon « internet live stats : le site international qui offre le nombre d'internautes connectés sur plusieurs réseaux sociaux, ainsi le nombre de postes publiée chaque jour », Le 13/04/2016 jusqu'à 14h, les statistiques montrent que [ref, i] :

- 3.347.880.432 utilisateurs d'Internet dans le monde.
- 139.721.178.702 emails envoyés.
- 388.716.620 tweets postés.
- 304.955.78 utilisateurs twitter.
- 1.640.411.904 utilisateurs Facebook connecté

En Algérie, avec la commercialisation de la 3G, entre janvier et juin 2014, la communauté algérienne sur Facebook a augmenté de 1 million, pour atteindre 7.8 millions d'utilisateurs en 2015 (selon Social Daily Statistics), positionnant ainsi Facebook en première position avec 96.59 % du nombre d'utilisateurs, avec une progression de deux points par rapport à 2013/2014, suivi de Twitter avec 1.28 % et Youtube (0.76 %).

Ces chiffres confirment que la population algérienne sont amateurs du réseau social Facebook, et sont moins séduit par les autres réseaux sociaux.
Après avoir présenté les médias sociaux, nous pouvons dire que leurs apparition, et plus parti-

culièrement les réseaux sociaux, ont donné une autre dimension à la notion de marketing et de communication, en permettant aux divers publics d'être de véritables acteurs de la communication, en offrant la liberté de partager leurs avis et leurs opinions sur tel ou tel produit. En effet, ces utilisateurs peuvent influencer d'autres par de simples messages et devenir des leaders d'opinion.

I.3 L'influence dans le web social

Le développement de l'internet permet plus que jamais aux utilisateurs de produire du contenu et d'entrer en interaction, surtout avec la montée des réseaux sociaux, la capacité de solliciter et partager des opinions et des idées à travers divers sujets ont subis des changements spectaculaires. Pour cela, la diffusion de l'information a longtemps été considérée comme un mécanisme important par lequel l'information peut atteindre de grandes populations et influencer l'opinion publique. Ceci a conduit les entreprises à déterminer une population cible qui a une capacité de diffusion et de médiatisation identifiable.
Cependant, étudier l'influence dans le web peut nous aider à mieux comprendre pourquoi certaines tendances ou innovations sont adoptées plus vite que les autres, et comment nous pourrions aider les annonceurs à concevoir des campagnes plus efficaces.

I.3.1 diffusion de l'information

Le phénomène de diffusion est observé et étudié depuis longtemps dans de nombreux domaines de la science, mais cela a commencé au début des années 50, quand [E and P.F, 1955] ont suggéré que les échanges avec l'entourage étaient plus persuasifs que les communications en provenance des entreprises via les médias de masse. Ainsi les entreprises auraient tout intérêt à communiquer auprès de certains individus appelés leaders d'opinion, qui propageraient ensuite le message commercial à leur entourage de manière plus efficace.
Ainsi l'information est diffusée en deux étapes, une petite minorité de « leaders d'opinion » (représentée sous forme d'étoile) agissent comme intermédiaire entre les médias et la majorité de la société (cercle).
Ces leaders d'opinion sont les plus exposés aux médias, se sont donc en grande partie eux qui filtrent, interprètent et transmettent les informations à leurs entourages. La **figure I.1** montre la diffusion de l'information selon [E and P.F, 1955].

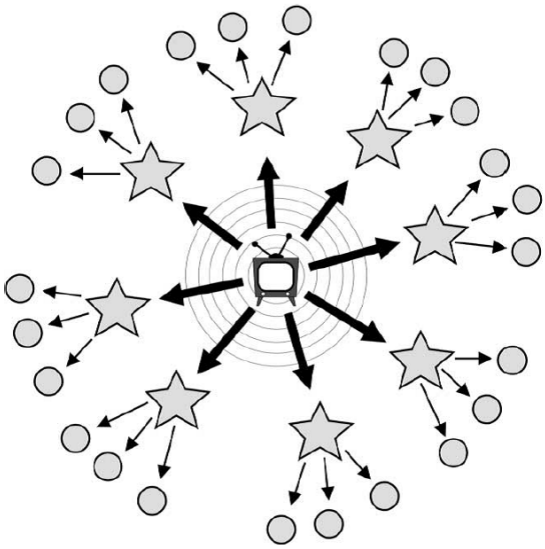

FIGURE I.1 – La diffusion de l'information

I.3.2 Concept d'influence

L'influence est un des mécanismes fondamentaux dont se préoccupe la psychologie sociale, elle montre à la fois l'emprise que la société exerce sur l'individu et les modifications qu'elles entraînent au niveau du comportement [G.N and C,].
Le terme d'influence désigne le processus par lequel une personne fait adopter une conduite ou un point de vue par une autre. L'influence sociale recouvre donc, tout ce qui produit un changement de la conduite suite à une relation ou une interaction avec un individu dit influent ou influenceur [ref, a].
À partir de ces définitions nous pouvons dire que l'influence se traduit par le fait que les actions d'un individu peuvent avoir un impact sur ses connexions et son entourage, en les induisant à se comporter d'une manière similaire. Son objectif est de jouer sur la perception et l'évaluation d'une certaine réalité en vue de les modifier.
Cependant, la notion d'influence est au cœur des recherches marketings, et surtout avec le développement du Web qui a engendré l'émergence de communautés virtuelles.Cette dernière est devenue très importante pour entreprises, car l'influence des utilisateurs sur leurs amis peuvent augmenter ou diminuer les ventes, pour cela les entreprises sont intéressés à trouver des personnes influentes et les encourager à créer une influence positive, ainsi déterminer quels sont les influenceurs clés pour leurs marques.

I.4 Détection d'influenceurs dans les réseaux sociaux

La notion d'influence joue un rôle essentiel dans le fonctionnement des entreprises, et de la façon dont une société fonctionne.
Ce qui fait qu'identifier l'influence sociale dans les réseaux est essentielle pour chaque entre-

prise,car elle leurs permet de comprendre comment les comportements se propagent. De plus, connaitre les mesures d'identification des influenceurs est nécessaire, afin de trouver les sources d'influence.

I.4.1 Les influenceurs et les leaders d'opinion

Un influenceur est une personne qui par son statut peut modifier les opinions de son public, résoudre les désaccords ou même de modifier leurs comportement de consommation. Connu comme un individu qui fait des contributions importantes de façon indirecte, son influence quant à elle peut être étendue, car il continue souvent à influencer un groupe même lorsqu'ils ne sont pas présents.Le secret de son efficacité réside dans la façon dont ils structurent ces demandes.
Cependant, L'étude [E and L, 2004] définit le leader d'opinion comme une personne qui exerce une force d'attraction (physique, psychologique et/ou sociale) sur son entourage et qui dispose d'une forte crédibilité dans une catégorie de produit. Ses jugements et comportements influencent les attitudes et les choix de marques de son entourage dans ce domaine.
Donc, les leaders d'opinion sont des individus clés suivi par un grand nombre de personnes, susceptible d'influencé un public vaste, grâce à la confiance que leur entourage l'ont attribué et la réputation d'analyseur pertinent qui se sont faite.
De ce fait, découvrir ces leaders d'opinions non seulement nous permettre de mieux comprendre les activités sociales qui se déroulent dans les médias, mais fournira également des opportunités uniques pour les ventes et la publicité.

I.4.2 L'impact des influenceurs sur l'image de l'entreprise

Les influenceurs sont devenus de plus en plus importants, car si ces derniers recommandent un produit, il mérite plus de confiance que la publicité traditionnelle, car selon un sondage de **Nielsen**, 85 % des consommateurs affirment prendre leurs décision après avoir consulté l'avis et les articles de ces nouveaux leaders d'opinion [ref, e].
La collaboration avec un influenceur et la communauté à laquelle il s'adresse, est une façon de partager une certaine expérience de la marque et permet d'établir une relation avantageuse pour toutes les parties. L'aide apportée par un influenceur lorsqu'il parle honnêtement d'un produit est un facteur qui contribue à capter l'attention du lecteur ou de l'utilisateur et peut même encourager l'achat, donc l'image de marque ne dépend plus exclusivement des messages émis par l'entreprise ou de ses campagnes de communication ; les consommateurs eux-mêmes contribuent à créer cette identité, par le biais de leurs expériences.
C'est pour cela que les entreprises cherchent à entrer en contact avec ce genre d'individus et être attentifs à ce qu'ils font, à ce qu'ils écrivent, mais toujours de manière sincère basés sur une confiance mutuelle, dans le but d'améliorer leurs images de marque et de promouvoir de nouveaux produits.
Mais parfois, ces influenceurs peuvent donner lieu à de nouvelles situations problématiques pour

les entreprises, une crise qui commence avec un avis d'un client mécontent, puis d'autres internautes qui renchérissent sur le sujet, ce qui engendre une chute pour la marque.

I.4.3 Le Réseau social : Twitter

Twitter est une plateforme de microblogging populaire créer en 2006 par **Jack Dorsey**, il permet aux utilisateurs de partager leurs opinions sur tous les domaines. Ce réseau a été évolué pour devenir un moyen pratique pour partager des opinions sur presque tous les aspects de la vie quotidienne. Par conséquent, les sites web de microblogging sont depuis devenus des sources de données riches pour l'analyse des sentiments.
La principale différence entre micro et traditionnelle blogs est la contrainte stricte de la taille du contenu, en effet Twitter permet à ses utilisateurs de publier des messages appelés "tweets", composé d'un maximum de 140 caractères. La limite imposer de caractères dans les tweets force les utilisateurs d'être concis dans leurs rédactions et d'être éventuellement, plus expressif et précis. De plus, en comparaison avec les autres réseaux sociaux, Twitter permet une grande variété d'usages, et propose une simplicité dans son interface, ce qui a contribué à son succès l'enregistrement de plus de 300 millions utilisateurs actifs, qui génèrent plus de 400 millions de tweets par jour. Par conséquent, Twitter occupe une place sans cesse plus importante dans notre environnement médiatique. Il est de fait devenu un outil de communication prisé par beaucoup de journalistes, acteurs de la vie politique ou encore des entreprises et des grandes marques.
En outre, Twitter permet d'accéder gratuitement à une part importante de ses données grâce à un simple API, ce qui pousse beaucoup de chercheurs à l'étudier et développer des recherches sur cette plateforme.

I.4.3.1 Fonctionnalités de Twitter

Le réseau social « Twitter » permet aux utilisateurs d'exprimer leurs avis dans n'importe quel domaine, avec des courts messages appelés « tweets », qui peuvent être « retweeté » par tout le monde vu qu'il est public.
Ce réseau se positionne par la relation sociale followers/following. La **figure I.2** illustre les différentes entités impliquées dans le réseau d'information et les diverses relations qui les associent.

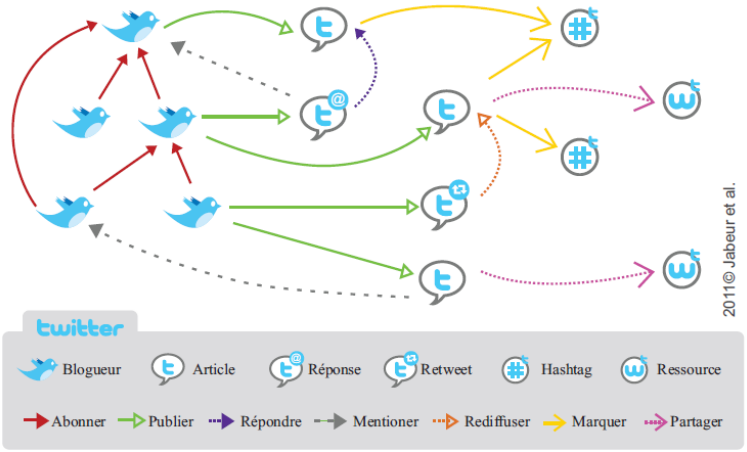

FIGURE I.2 – Le réseau d'information Twitter

Les tweets publiés sont en principe des textes, mais peuvent également contenir des images, des vidéos, des URL, ainsi des hashtags. Dans le **Tableau I.1** suivant nous montrons les différentes fonctionnalités.

Fonctionnalité	Description
Tweet	Le tweet est un message court posté par un utilisateur dans le but de donner un avis ou un sentiment sur un sujet donné.
Retweet	le Retweet se présente sous forme d'un tweet republié à partir d'un tweet d'un autre utilisateur, précédé par l'abréviation RT.
Following / Abonnements	C'est le nombre de comptes Twitter que suit un utilisateur. Ce nombre varie d'une personne à une autre selon les tendances de l'utilisateur.
Followers / Abonnés	c'est le nombre de comptes Twitter qui suit cette personne. Ce nombre varie d'une personne à une autre selon la popularité d'un utilisateur.
Hashtags	un hashtag est un mot précédée du symbole (#), ce mot représente généralement un mot clé du message posté, il est utilisé pour catégorisé les messages sur un sujet précis afin de facilité la recherche sur twitter.
Mentions	Une mention est un nom précédé d'un @, elle est utilisé pour spécifié un utilisateur donné dans un post (adressé le message posté), cette mention représente un lien directe vers l'utilisateur, ce dernier est informé automatiquement par une notification.
Timeline	La timeline est la page principale sur laquelle apparaissent le fil d'actualité qui représente les tweets des abonnés d'un utilisateur.
TrendingTopic	Les tendingtopics sont les sujets tendance sur twitter à un moment donné, pour un pays donné, voire tous les pays confondus.

TABLE I.1 – Fonctionnalités de Twitter

I.4.3.2 Structure de données d'un utilisateur Twitter

Comme chaque média social, l'utilisateur doit avoir un compte contenant ces informations nécessaires, afin de pouvoir interagir avec les autres. Dans ce qui suit, nous allons détailler les informations les plus importantes d'un compte Twitter :

- **Name :** Le nom de l'utilisateur, choisi par lui-même. Pas nécessairement le nom d'une personne. Ne dépasse pas les 20 caractères, mais sous réserve de modification.
- **Screenname :** c'est le pseudonyme que cet utilisateur s'identifie, doit être unique, mais sous réserve de modifications. Typiquement, un maximum de 15 caractères. Il est toujours suivi

par @ (@screenname).

- **Date de création :** la date de création du compte de l'utilisateur sur Twitter.
- **Location :** c'est un champ non nul, qui représente l'emplacement défini par l'utilisateur pour le profil de ce compte.
- **statuses_count :** c'est le nombre de tweets (y compris les retweets) émis par l'utilisateur.
- **Relation d'un utilisateur :** c'est le nombre de followers et de following de l'utilisateur.

I.4.3.3 L'influence sur Twitter

L'influence sociale a été définit comme étant « un phénomène social que les utilisateurs des médias sociaux peuvent subir et exercer, traduisant le fait que les actions d'un utilisateur peuvent induire ses connexions à se comporter d'une manière similaire. L'influence se manifeste parfois explicitement dans les médias sociaux, par exemple sur Twitter, lorsqu'un utilisateur retweet » [A, 2014].

En effet, le but initial de Twitter était de proposer un service permettant à ces utilisateurs de tenir en permanence leur entourage de leurs faits et gestes. Cependant, l'utilisation a considérablement évolué et permet à tout individu d'interagir avec ces followers à travers leurs tweets, et leur donne la possibilité de diffuser des informations sur n'importe quel sujet et partager son avis sur l'actualité.
Twitter est également un service de réseau social : pour lire les « tweets » d'un individu il faut le suivre, c'est-à-dire s'inscrire sur le réseau Twitter et s'abonner au flux que propose l'utilisateur, sans la nécessité de l'accord de l'émetteur des « tweets » pour les lire. Ce qui fait que l'information peut se propager rapidement, ainsi les individus peuvent influencer les autres rapidement.
D'après [L et al., 2011] , l'influence d'un utilisateur dépend de ses relations de rediffusion et elle est estimée selon sa position dans le réseau.

I.4.4 Mesures de détection d'influenceurs sur twitter

La détection d'influenceurs est devenue primordial pour les entreprises, car un influenceur peut les aider à améliorer l'image ou la notoriété de la marque, comme il peut la détruire. Cependant, mesurer l'influence sur Twitter dépend de plusieurs critères :

- **La popularité (Le nombre de followers) :** pour qu'un individu soit un influenceur, il faut qu'il ait quelqu'un à influencer. Donc le nombre d'abonnés parait être un candidat potentiel au titre d'indicateur du leadership. Il est même un critère souvent cité par les professionnels pour signifier l'influence d'un individu [T, 2009].
  A ce sujet [Augure,] a classer les utilisateurs selon leur popularité en trois catégories, comme le montre la pyramide d'influence dans la **figure I.3** suivante :

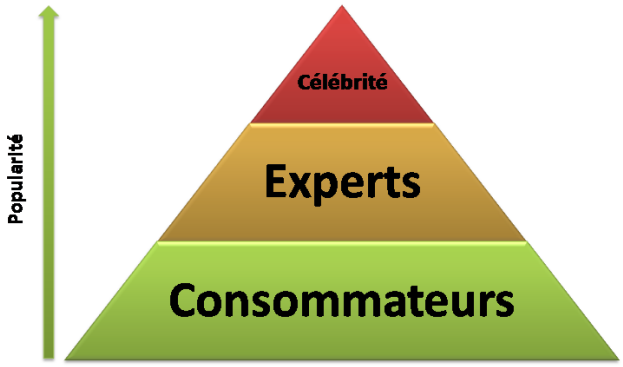

FIGURE I.3 – La pyramide de l'influence

Dans ce qui suit, nous allons détailler chaque palier de la pyramide [Augure,] :

– **Catégorie 1 : Célébrité**

  Dans cette catégorie, nous retrouvons les fameuses « stars ». Elles exercent une influence « globale » ; peu importe les sujets sur lesquels elles s'expriment, cela aura de l'impact sur leurs fans. Ceux sont des utilisateurs ayant un compte certifié avec plus de 40 milles personnes dans leurs cercles d'amis. L'intérêt principal des comptes certifiés est d'éviter les « faux » comptes, et ainsi d'empêcher l'usurpation d'identité des célébrités et des personnes connues.

– **Catégorie 2 : Experts**

  Il peut s'agir de journalistes connectés, de blogueurs, de dirigeants du monde de l'entreprise ou du monde associatif, d'activistes, d'experts de leur secteurs, … etc, ils ont des profils variés mais ont tous un attribut en commun, ils sont crédibles vis-à-vis de leurs communautés et sur des thématiques précises : un message de leur part suscitera un vif intérêt aux yeux de leurs suiveurs qui partagent leurs centres d'intérêt et accordent beaucoup de crédit à leurs opinions.

– **Catégorie 3 : Consommateurs**

  Cette catégorie a pris beaucoup d'importance ces dernières années avec l'avènement des réseaux sociaux. Les consommateurs représentent le grand public. Ils peuvent facilement devenir de véritables ambassadeurs d'une marque ou, au contraire, des détracteurs dont l'impact peut avoir un effet dévastateur.

  En 2013, un passager mécontent de la compagnie aérienne British Airways a fait beaucoup parler de lui avec son tweet de mécontentement. Ce dernier a été repris dans de nombreux médias et a même été interviewé par la chaîne CNN. (Fille mm que site précédent)

- **L'impact (Le ratio abonnés/abonnements) :** le calcul du ratio du nombre d'abonnés par rapport au nombre d'abonnements a été définit comme étant un indicateur d'influence, car plusieurs utilisateurs suivent beaucoup de gens, alors qu'ils ne sont pas suivis. Donc un influenceur doit avoir un rapport abonnés/abonnements élevé.
- **Nombre de retweet :** Une fonctionnalité spécifique à Twitter est peut être considéré comme indicateur d'influence, en effet, un influenceur aurait son contenu plus fréquemment répété et rediffusé par la communauté [E et al., 2012].

I.4.5 Les travaux reliés

De nombreux acteurs de la société, tel que les entreprises cherchent à exploiter et analyser les médias sociaux à des fins diverses (analyser la réaction des consommateurs à propos de certains produits et les promouvoir, détecter des informations et utilisateurs dangereux, détecter des évènements importants et interroger les utilisateurs). Généralement, la démarche qu'ils cherchent à mettre en œuvre consiste à détecter les évènements animant les discussions des utilisateurs, puis à identifier les utilisateurs influents par rapport à ces évènements, afin de prendre des décisions et éventuellement agir.
De ce fait, plusieurs travaux ont été proposés pour répondre à ce besoin, tel que[J.P and A, 2010] qui ont déduit que le nombre de suiveurs n'est pas pertinent comme indicateur du leadership de l'utilisateur, car une certaine proportion d'individus suivent en retour ceux qui les suivent, créant ainsi une relation entre le nombre de suivis et le nombre de suiveurs, ceci augmente l'opportunité de faire augmenter le nombre de suiveurs. De ce fait, ils suggèrent que le ratio « suiveurs / suivis » est un indicateur pertinent qui permet de détecter les vrais influenceurs. Ce ratio à l'avantage d'être facilement accessible (le nombre de suiveurs et de suivis étant affichés sur le profil de chaque utilisateur) et utilisable sans délai par les entreprises.
Cependant, ce ratio peut donner des résultats douteux, dans le cas où un utilisateur à un nombre de suiveurs et suivis réduit.

I.5 Conclusion

Dans ce chapitre, nous avons expliqué le marketing viral pour les entreprises, la naissance des médias sociaux et leur importance, ainsi les concepts d'influence et des influenceurs.
De plus, nous avons présenté le réseau Twitter, qui représente le réseau social sur lequel notre approche a été définie et ces différentes fonctionnalités. De même, nous avons parlé de l'influence sur Twitter et comment on peut mesurer et identifier les influenceurs. Cependant, les entreprises ont intérêt à connaître ces influenceurs, et cela tout en analysant le contenu de leurs publications.
De ce fait, dans le chapitre suivant, nous allons présenter comment l'information se propage et détailler le concept de fouille d'opinion.

CHAPITRE II

L'analyse des sentiments pour l'identification d'influenceurs dans les Réseaux Sociaux.

A l'origine, l'information était le monopole des médias, de la presse écrite, du journal télévisé . . . etc, mais depuis l'émergence d'internet, et l'apparition des réseaux sociaux la toile s'enflamme avec un grand nombre d'utilisateurs connecté suscitent leurs intérêt à divers événements et informations,s'engageant ainsi fortement dans cette nouvelle technologie qui garantit un accès simple et rapide, et une gratuité totale lui offre une possibilité de partager et de donner son avis, informer le public ou même critiquer un produit.
De ce fait, l'analyse des sentiments est primordiale pour déterminer l'opinion, l'émotion ou le jugement d'une personne concernant un sujet spécifique, qui peut offrir des avantages à une variété de domaines, à partir des prévisions de vente, à la politique et les investisseurs.
Le but de cette analyse est de connaitre l'opinion des utilisateurs, à partir du contenu de leurs publications, afin d'identifier les influenceurs.
Dans ce second chapitre, Nous commençons par présenter la propagation de l'information, le principe de la recherche d'information, ainsi la recherche d'information sociale. Puis nous parlerons de la fouille d'opinion et les différents travaux reliés.

II.1 Propagation d'information

Dans une société où l'utilisation des médias sociaux est devenue un quotidien de la vie, la propagation de l'information est devenue immédiate modifiant ainsi les rôles d'auteur, éditeur et lecteur, mettant l'utilisateur au centre de l'univers informatif. Cette avancée technologique à modifier considérablement nos habitudes de consommation, poussons l'usager à se connecté pour diffuser de l'information, connaitre les nouveautés ou échanger les avis avec son entourage.
Cette nouvelle façon de diffuser l'information en temps réel augmente la productivité des entreprises, en leurs permettant de fournir et de partager davantage d'informations au prés de leurs clients rapidement et pour un faible coût de possession, tout en déplaçant l'information d'un nœud à un autre du réseau, d'une communauté à une autre provoquant une masse importante de données.
De ce fait, les médias sociaux sont devenus des outils très puissants de propagation,faisant d'eux des sources d'informations précieuses pour les divers marques et entreprises, qui veulent soigner leurs images et leurs présences en ligne, ce qui a engendré les chercheurs à augmenter leurs travaux

sur la recherche d'information à travers ces médias.

II.2 La recherche d'information

Les récents progrès des technologies de l'information de manière générale et des réseaux de communication de manière particulière, ont redonné à l'information de nouveaux contours et davantage de valeur selon divers aspects : scientifique, technique, économique, d'usage ... etc.
Actuellement, l'information est devenue notre matière première, mais il faudrait savoir la localiser et la sélectionner, ce qui fait que la Recherche d'information (RI) est un domaine qui s'intéresse à la structure, à l'analyse, à l'organisation, au stockage, à la recherche et à la découverte de l'information.
Cependant, [N, 2006] à définit la Recherche d'information (RI) comme étant « une activité dont la finalité est de mettre en regard des informations et un utilisateur. C'est une activité par laquelle un utilisateur accède à un granule d'informations (un ensemble de documents, un document, une partie de document, un composant XML, une donnée) à partir "un besoin qu'il spécifie ».
Avec l'explosion des technologies web et la naissance des réseaux sociaux, la recherche d'information sociale est devenue nécessaire.

II.3 La recherche d'information sociale (RIS)

Depuis la création d'internet, le web a connu un succès gigantesque et est devenu peu à peu le premier outil pour la production, la publication, la diffusion et le partage de l'information. De plus, Les technologies du Web 2.0 mettent l'utilisateur au centre de la production de données et introduisent une forte composante collaborative et sociale.
Les statistiques récentes montrent la montée en puissance d'utilisation des réseaux sociaux, selon les dernières mises à jour de [ref, g] le nombre d'utilisateurs de Facebook seulement est 1.59 milliard, suivi par YouTube et WhatsApp avec 1 milliard d'utilisateurs.
la figure II.1 montre le nombre d'utilisateurs actifs des réseaux sociaux dans le monde.

FIGURE II.1 – Nombre d'utilisateurs actifs sur les réseaux sociaux dans le monde

II.3.1 La recherche d'information sociale sur Twitter

La recherche des tweets est une tâche de recherche d'information ad-hoc dont l'objectif est de sélectionner les tweets pertinents en réponse à une requête Q [I et al., 2011].
Quand on parle de pertinence dans la recherche de tweets, on ne se limite pas à la similarité textuelle mais nous prenons également en compte les interactions sociales dans le réseau. De ce fait, la pertinence des tweets dépend aussi de l'importance de l'utilisateur qui les publie [L et al., 2011].
En comparaison avec une recherche classique sur le web, la recherche sur twitter est plus précise vu la limitation du tweet. Elle permet aussi de recevoir en temps réel des informations sur un événement qui vient de se produire quelques secondes auparavant.La recherche de tweet permet également d'accéder aux actualités avec une diversité de points de vue des utilisateurs et à une échéance proche de l'événement [L et al., 2011].
En utilisant l'API de Twitter, la recherche des tweets devient plus facile. Selon le type de l'API utilisée, les résultats de recherche seront axés soit sur la pertinence soit sur l'exhaustivité. Ainsi, la recherche s'effectue par mots-clés, et cela suivant la requête de l'utilisateur. La requête peut contenir des opérateurs qui modifient les résultats.
L'étude [L et al., 2011] associe la pertinence des tweets à l'importance sociale de l'utilisateur. Elle considère l'influence et l'expertise comme les principaux facteurs sociaux, qui déterminent l'importance de l'utilisateur et la qualité de ses tweets.

II.4 Fouille d'opinion

Avec l'essor du web, les gens donnent de plus en plus, leurs avis sur internet à propos d'un grand nombre de sujets, ainsi la manière dont les personnes expriment leurs opinions a beaucoup changé, et l'information est devenu à la portée de tout le monde, incitant les internautes à suivre ces avis avant leurs prises de décision et d'exprimer en retour leurs opinions à ce sujet.
Les différentes organisations se sont rendu compte que l'opinion exprimée constitue une importante source d'information, et qu'il faut prévoir de suivre en temps réel ces données stratégiques et d'élaborer des techniques de traitement automatique, afin de transformer ces données en connaissances utiles et exploiter intelligemment cette source.
Dans notre travail, nous nous intéressons en particulier à l'analyse des données liées à l'opinion ou aux sentiments exprimés par les internautes au sein des réseaux sociaux.

II.4.1 Définition

Fouille d'opinion, aussi appelée analyse des sentiments, été définie comme « un sous-domaine de la fouille de textes qui consiste à analyser des textes, afin d'en extraire des informations liées aux opinions et sentiments ».[D, 2011]

[K, 2013] quant à elle a définit l'analyse des sentiments comme « l'étude de calcul des opinions, des sentiments, la subjectivité, les évaluations, exprimée dans le texte ».
A partir de ces deux définitions nous pouvons dire que la fouille d'opinion concerne l'extraction d'un sentiment dans une source, telle qu'un texte sans structure prédéfinie, avec une classification du sentiment, suivant l'opinion générale exprimé, c'est-à-dire une classification en deux classe (positif ou négatif) , ou en trois classes (positifs , négatif ou neutre), ou en des classes définies plus finement.

Avec l'ampleur que prennent les réseaux sociaux, l'analyse des sentiments est devenue plus particulièrement utilisée dans le domaine du marketing, pour extraire l'opinion que peut exprimer les internautes dans leurs différentes postes publiées sur leurs produits ou services.Cependant,la fouille d'opinion rencontre certaines difficultés que nous citrons dans ce qui suit.

II.4.2 Le problème de fouille d'opinions

L'étude des opinions et sentiments dans les textes est un axe de recherche qui s'est constitué au début des années 2000.
L'opinion peut-être définie selon [B, 2010] comme « l'expression des sentiments d'une personne envers une entité, étant subjective et opposée à l'expression de faits ».
Dans l'exemple de la **figure II.2**, certaines phrases (1, 6) sont simplement factuelles (ou objectives), tandis que d'autres (2, 3, 4, 5) relatent le sentiment du critique envers l'entité (subjectives).

"(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) Although the battery life was not long, that is ok for me. (6) However, my mother was mad with me as I did not tell her before I bought it. ... "

FIGURE II.2 – Extrait d'évaluation de l'iPhone [B, 2010]

Néanmoins la fouille d'opinion met en évidence d'autres problèmes techniques causés par certains facteurs, nous citons :

- **Hétérogénéité des données :** Les données d'opinion récupérée sont écrites dans un langage plus au moins familier, contenant des phrases qui utilisent des abréviations rendant l'extraction difficile, et empêchent l'analyse correcte de ces opinions.
- **Dépendance du contexte :** En général, les sentiments sont très sensibles au contexte, nous pouvons trouver des mots qui peuvent indiquer une opinion positive ou une opinion négative sur deux aspects différents, par exemple, on peut exprimer une opinion négative avec des mots donnant une opinion positif, comme suit : « j'aime bien l'iPhone qui marque batterie faible, alors qu'il est en train de charger depuis une heure ».
- **Dépendance du domaine :** Le Problème survient du changement de vocabulaire, par exemple, la même expression peut indiquer différents sentiments dans différents domaines.
  Les différentes approches évoqués dans la littérature ont été basés sur le lexique du texte, ou par l'apprentissage, afin d'analyser le sentiment du critique, et leurs efficacité a été confirmée.

II.4.3 Les approches de la fouille d'opinion

L'analyse des sentiments est considérée comme source d'innovation dans le domaine de la classification de texte, c'est pour cela que plusieurs chercheurs s'intéressent à ce domaine-là. Généralement, il existe deux méthodes différentes afin de procéder à l'analyse automatique des sentiments, la première méthode est basée sur un lexique, construit à partir de dictionnaire existant, la deuxième méthode est basée sur le corpus comportant des textes évaluatifs, dont le langage est généralement subjectif.

II.4.3.1 Approche basée sur le lexique

La plupart des études de l'analyse des sentiments se sont basées sur des lexiques. Ceux-ci permettent d'évaluer la polarité d'un texte subjectif à l'aide de deux groupes de mots : ceux qui expriment un sentiment positif, et ceux qui expriment un sentiment négatif. Le système de traitement extrait du texte tous les mots positifs et négatifs, il s'agit pour la plupart des verbes (aimer,

apprécier, détester, ...) et d'adjectifs (magnifique, insupportable, ...), mais aussi de quelques noms communs (plaisir) et d'adverbes (malheureusement). De plus, il y a des règles plus complexes pour extraire les relations des phrases plus compliquées. L'attribut de la relation (positif ou négatif) d'un sentiment sera inversé quand une négation est présente dans la phrase, comme par exemple : « Je n'aime pas le nouveau design de l'iPhone» « Ce n'est pas un mauvais smartphone. ». Si la quantité de mots positifs l'emporte sur celle de mots négatifs, le logiciel tend à dire que le texte exprime un sentiment négatif. Inversement, le texte est considéré comme positif.

II.4.3.2 Approche par apprentissage

Contrairement à la méthode basée sur un lexique, l'analyse automatique des sentiments basée sur un apprentissage ne s'appuie pas sur un vocabulaire de mots positifs et négatifs. En revanche, elle requiert l'élaboration de deux corpus annotés manuellement. Le premier corpus constitue le corpus d'apprentissage qui s'utilise afin d'entraîner un système automatique.
Il comporte des notes ajoutées par des annotateurs humains. A partir de ces notes, le système automatique devrait être capable de procéder à une analyse pareille de façon autonome. Le deuxième corpus constitue le corpus de test, celui-ci est élaboré afin de vérifier la performance du système automatique. Dans un scénario idéal, les résultats de l'analyse faite par le système automatique correspondraient cent pour cent avec ceux du corpus d'apprentissage. Afin que la performance du système automatique soit maximale, il importe que le corpus d'apprentissage soit représentatif pour le corpus de test.
Il existe deux méthodes d'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé. Dans La méthode « supervisée » le système automatique est entraîné à traiter une base de données divisé dès le début en groupe. Par contre, la méthode « non supervisée » le système doit lui-même structuré les informations au sein du corpus en les divisant en groupes. Ainsi, il doit organiser la base de données d'une telle manière que les données les plus similaires soient associées dans un groupe et les données différentes dans un autre.

II.4.4 Les travaux réalisés sur l'analyse des sentiments dans les réseaux sociaux

Avec la naissance d'internet, et plus particulièrement les réseaux sociaux, des millions d'utilisateurs partagent des opinions sur différents aspects de la vie quotidienne, ce qui a amené les chercheurs à augmenter leurs efforts pour effectuer l'analyse des sentiments.
Comme nous l'avons déjà vu, avec l'analyse du sentiment, nous essayons d'extraire l'émotion ou de l'attitude d'un document ou d'un morceau de celui-ci. Dans les réseaux sociaux un document pourrait être un tweet ou un post.

II.4.4.1 Approches basée sur les émoticônes

Plusieurs chercheurs ont appliquées des méthodes basé sur le vocabulaire des émoticônes, tel que l'approche [A and P, 2010], où ils ont recueillis un corpus de messages texte et formé un en-

semble de données de trois classes : sentiments positifs, sentiments négatifs, et un ensemble de textes objectifs (pas de sentiments).
Pour recueillir des sentiments négatifs et positifs, ils ont interrogé Twitter pour deux types d'émoticônes : des émoticônes heureux : " :-)", " :)", "=)", " : D" etc, et des émoticônes triste : " :-(", " :(", "= (", " ; (" etc.où ils ont supposé qu'une émoticône dans un message représente l'émotion que porte ce message et que tous mots du post sont liés à cette émotion.
Les deux types de corpus recueillis ont été utilisés pour former un classificateur pour reconnaître des sentiments positifs et négatifs, en utilisant Naive Bayes, MaxEnt et Support Vector Machines (SVM).
Cependant, les résultats de ces recherches ne sont pas précisent, et les individus expriment généralement leurs sentiments sans pour autant utiliser des émoticons.

Semblablement, on trouve que [A et al., 2011] propose dans leur travail deux ressources pour le prétraitement des données : un dictionnaire émoticône et un dictionnaire acronyme, où ils ont préparé le dictionnaire d'émoticône par marquage de 170 émoticônes figurant sur Wikipédia avec leurs état émotionnel, alors que le dictionnaire d'acronyme contient 5184 acronymes. Le corpus recueillis a été soumis à l'étape de prétraitement qui consiste à remplacer tous les émoticônes avec leur polarité de sentiment, en consultant le dictionnaire d'émoticône et les acronymes par leurs expansions, dans le but de classer les termes de chaque publication selon leur polarité respective.
Une fois les données récoltées, les deux travaux s'intéressent à une phase de classification. Les méthodes les plus utilisées pour la classification d'opinions sont les méthodes de classification supervisée. Ce type de méthodes consiste d'abord à construire un corpus d'apprentissage (classification à l'aide d'exemples), ces derniers sont des données dont on connait déjà la classe. On parle dans ce cas de données classées ou étiquetées. Ensuite, un classificateur est élu pour prédire avec précision la classe cible pour chaque cas dans les données. Beaucoup de méthodes de classification supervisée existent et beaucoup d'entre elles ont été testées pour la classification d'opinions.
Toutefois, les méthodes les plus présentes dans la littérature, et qui semblent également être les plus performantes sur les textes, sont les Machine à Vecteurs de Support (SVM) [A, 2007] et les classificateurs Naïfs Bayésiens(NB). [B and L, 2004]

Néanmoins, les résultats générer sont douteux, car certaines publications sont classées dans des catégories incongru suite à une utilisation incorrecte des émoticônes dans les postes.

II.4.4.2 Approche basée sur le lexique

De nombreuses méthodes automatiques de fouille d'opinion s'appuient sur un lexique, dans lequel à chaque entrée est associé un degré de polarité. La construction de telles ressources linguistiques est donc devenue un champ de recherche important en linguistique computationnelle (informatique).

En effet, ces approches dépend de l'opinion (ou sentiment) des mots, où le texte est considéré

comme un ensemble de mots souvent sans structure (représentation « sac de mots »). Les mots d'opinion sont généralement contenues dans un dictionnaire appelé l'opinion lexique.

Comme [M et al., 2011] qui est basée sur des mots d'opinion, à savoir, les mots qui sont couramment utilisés dans l'expression positive ou négative regroupés dans un sac avec un score, qui représente le taux de polarité de chaque mot. Ensuite, pour chaque message donné, ils ont remédié à une extraction des mots porteurs de sentiments (y compris les adjectifs, les verbes, noms et adverbes), et les annotes avec leurs valeurs de polarité, en se référant au dictionnaire, afin de calculer l'orientation sémantique du texte, en tenant compte de la négation.
Après avoir collecté les données, une classification est faite de la même façon que les approches basé sur les émoticônes.

Seulement, l'utilisation de « sac de mots » ne se révèle pas toujours efficace, car les données issues du langage naturel sont séquentielles, et la recherche ne peut jamais être générale vue le nombre de mots contenues dans le sac.

II.4.4.3 Approche basé sur une ontologie

La récolte de données cités précédemment ne donnent pas le résultat profond de l'opinion des utilisateurs, car ses dernières ont été faites sans extraction de caractéristiques, or on peut trouver des publications qui expriment deux opinions différentes sur des caractéristiques différentes de la même marque, par exemple : « la batterie de sonyXperia était merveilleuse, bien que la taille de l'écran était mauvaise ».
De ce fait, [K and D, 2015] ont proposés une technique, où l'idée de base derrière l'approche proposée consiste à tirer parti d'une ontologie de domaine pour fournir des scores de sentiment plus élaborées concernant les notions contenues dans un tweet.
L'objectif est d'avoir un système qui accepte en entrée un tweet (ou un ensemble de tweets) sur un sujet spécifique et fournit des scores de sentiment pour chaque aspect / caractéristique de ce sujet.
La méthodologie proposée est divisé en deux phases : (a) la création de l'ontologie de domaine, et (b) l'analyse des sentiments sur un ensemble de tweets, basée sur les concepts et les propriétés incluses dans l'ontologie.

a) Création de l'ontologie de domaine Afin de créer une ontologie de domaine, ils ont opté pour deux méthodes, la première méthode est l'Analyse formelle de concept, et la deuxième méthode c'est l'apprentissage de l'ontologie.

- Analyse Formelle de Concept(AFC) Analyse Formelle de Concept est une théorie de l'analyse des données mathématiques, généralement utilisé dans la représentation des connaissances et gestion de l'information. Sa principale caractéristique est qu'il applique une méthode étape par étape axée sur l'utilisateur pour créer des modèles de domaine. Avec l'émergence récente du Web sémantique, selon [K and D, 2015] l'analyse Formelle de Concept a été comptabilisée comme un outil d'ingénierie précieux

pour dériver une ontologie à partir d'une collection d'objets et de leurs propriétés.
Cette théorie a été préférée dans leur travail, car il offre les avantages suivants :

- Appropriée la taille de l'ontologie.
- Une meilleure conception de l'ontologie.
- Domaine ontologie spécifique.

Algorithm *createOntology*

Input: A default concept (c)
Variables: An initially empty set of tweets (W)
An initially empty set of objects (O)
An initially empty set of attributes (A)
Output: A cross-table, filled with objects and attributes (T)

```
1.  W ← retrieveTweets(c); // automatically choose the first n tweets
2.  foreach w ∈ W do
3.      o ← retrieveObject(w);
4.      if o ≠ NULL then
5.          O := O ∪ {o};
6.          A' ← retrieveAttributes(w);
7.          foreach a ∈ A' such that (o, a) ≠ ∅ do
8.              A := A ∪ {a};
9.  T ← populateTable(O, A);
10. return T
```

FIGURE II.3 – Algorithme de la création de l'ontologie par la AFC [K and D, 2015]

- Apprentissage de l'ontologie Dans cette partie, les auteurs ont eu recours à OntoGen qui est un éditeur semi-automatique piloté par les données de l'ontologie, afin de faire un apprentissage avec l'ensemble de tweets récupérés.
Après avoir créé l'ontologie via AFC ou l'apprentissage de l'ontologie, ils ont enrichi l'ontologie avec des synonymes et hyponymes, en utilisant la base de données Wordnet.

b) l'analyse des sentiments Après avoir collecté un ensemble de tweets, basée sur les concepts et les propriétés incluses dans l'ontologie, les tweets récupérés sont soumis à OpenDover pour l'analyse de sentiment. OpenDover est un service web qui tags les opinions et les sentiments détectés dans un corpus textuel, basé sur un domaine. Or, OpenDover a été considéré comme un choix approprié pour leur approche, car il est adapté pour extraire le sentiment des phrases isolées.

II.4.5 Méthodes de classification

Les médias sociaux regorgent d'informations dans différents domaines de la recherche sociale, les données récupéré sont riches en matière de prise de décision.

Comme il a déjà été précisé dans les différents travaux cité dans la section précédente, c'est principalement la classification automatique du texte qui se charge d'affecter pour chaque information une catégorie en fonction de son contenu.
Dans cette optique, il existe deux grandes familles de classificateurs regroupent plusieurs algorithmes misent au point pour des problèmes quelconques en apprentissage automatique.

II.4.5.1 Types de classification

La classification automatique est une catégorisation qui permet d'attribuer à un objet (individu) une classe ou catégorie. Cette classification se présente sous deux formes :

a) Classification supervisé La classification supervisée est utilisée lorsqu'on possède un ensemble de classes prédéfini, et qu'on devrait à partir des caractéristiques d'un nouvel objet l'affecter à une classe de cet ensemble. Ce type de méthodes consiste à construire un modèle de classification à l'aide d'exemples. Des exemples sont des données dont on connait déjà la classe. On parle dans ce cas de données classées ou étiquetées, à partir de là un classificateur est générer obtenant ainsi un modèle de prédiction qui permet par la suite de prédire la classe d'un objet nouvellement ajouté.

b) Classification non supervisé La classification non supervisée est utilisée lorsqu'on possède des objets qui ne sont pas classées et dont on ne connait pas la classification. A la fin du processus de classification, les objets doivent appartenir à l'une des classes générées par la classification, on distingue deux catégories de classification non supervisé : hiérarchique et non hiérarchique.

II.4.6 Algorithmes de classification

Il existe dans la littérature un grand nombre d'algorithmes de classification, qui partagent tous le but d'assigné une catégorie au données faisons l'objet de classification.
Au niveau de cette partie, nous allons aborder quelques algorithmes de classification connus pour être sollicité dans le domaine de classification automatique du texte.

a) Machine à support de vecteurs (SVM) Le principe des SVM consiste en une stratégie de minimisation structurelle du risque, cet algorithme permet de déterminer si un élément appartient (qualité de positif) ou non (qualité négatif) à une classe, l'idée est de créer par apprentissage une surface de séparation entre des exemples positifs et négatifs, minimisant le risque d'erreur et maximisant la marge entre deux classes grâce à un Hyperlan, qui sépare les documents appartenant à la catégorie et ceux qui n'ont fait pas partie [S, 2005].

b) Naïve bayés L'algorithme naïve bayés se base sur le théorème de Bayes, permettant de calculer les probabilités conditionnelles, c'est-à-dire, on cherche la classification qui maximise la probabilité d'observer les mots du document. Il procède tout d'abord à la phase d'entrainement, en calculant la probabilité que ce mot soit présent dans un texte de cette même

catégorie. Par la suite, quand un nouveau document doit être classé, on calcule les probabilités qu'il appartient à chacune des catégories à l'aide de la règle de Bayes [T and M, 1997].

c) K plus proche voisin (KNN) L'algorithme K –voisins les plus proches (KNN) figure parmi les plus simples algorithmes de classification automatique, dans un contexte où un nouveau texte x à classer arrive, l'idée fondatrice simple est de comparé chaque nouveau texte x à l'ensemble des textes du jeu d'apprentissage. La classe de x est déterminer en fonction de la classe majoritaire parmi les k plus proches voisins, soit en moyenne celle qui contient le plus de textes voisins

II.5 Conclusion

Dans ce chapitre, nous avons parlé dans un premier lieu de la propagation d'information, de la recherche d'information, et la recherche d'information sociale.
Dans un second lieu, nous avons présenté brièvement le domaine de la fouille d'opinion, du problème de classification d'opinions, qui est l'un des sujets les plus étudié dans ce domaine, ainsi les différentes techniques et algorithmes utilisés pour la collection et la classification des données.
Puis, nous avons présentés les travaux les plus pertinents pour l'analyse de sentiment dans les réseaux sociaux.
Dans la prochaine partie du mémoire, nous allons détailler notre approche pour la recommandation d'influenceurs dans les réseaux sociaux, en se basant sur l'analyse du contenu des publications.

Deuxième partie

Notre Contribution

CHAPITRE I

Conception de l'approche

L'influence a longtemps été étudiée dans les domaines de la communication et de marketing, surtout avec l'apparition des réseaux sociaux, du fait qu'elle joue un rôle essentiel dans le fonctionnement des entreprises, car elle peut aider à développer leur image, comme elle peut entrainer une chute de l'entreprise dans le cas d'une influence négative.

De ce fait, étudier l'influence aide à mieux comprendre pourquoi certaines tendances ou innovations sont adoptées plus vite que les autres, et comment nous pourrions aider les annonceurs à concevoir des campagnes de marketing plus efficaces. De plus, en ciblant les personnes influentes dans le réseau, permettra aux entreprises à mieux gérer leur image en contactant ses derniers, et cela avec un coût de marketing très petit.

Ainsi, si nous devons schématiser ce problème : toute entreprise souhaitant augmenter la commercialisation de ses produits, doit trouver un sous ensemble d'individus, dit influenceurs. Ces derniers pourront aider à améliorer l'image de la marque par leurs critiques ou de promouvoir leurs produits.C'est pour cela, détecter et identifier ses influenceurs est devenu une nécessité pour toute entreprise.

Dans ce chapitre, nous présentons notre approche pour la détection des influenceurs, en se basant sur l'analyse du contenu des publications issues des réseaux sociaux. Dans le cadre de notre travail, nous avons choisi de centrer notre approche sur le réseau social Twitter, car ce dernier est très utiliser pour les campagnes de marketing, et il est devenu un outil de communication très populaire pour un pourcentage considérable d'utilisateurs, et un moyen raisonnable pour le partage des opinions sur tous les aspects de la vie. De plus, en raison de la limite de 140 caractères par message, il est considéré comme une source de données pour l'exploitation de l'analyse des sentiments.

Notre approche consiste d'abord à construire un corpus d'étude via des mots clés, à l'aide d'un API Twitter et une ontologie pour l'extraction des caractéristiques du domaine. Les données récupérées nécessitent un élagage pour avoir que des données pertinentes, puis une catégorisation des données par caractéristique aura lieu. Une fois le corpus d'étude soit prêt,on passe à l'étape d'analyse de sentiment qui consiste à attribuer la polarité adéquate pour chaque post, à l'aide d'une classification basée sur un apprentissage supervisé.
On commence le présent chapitre par le méta modèle du réseau social étudié, puis on enchaine par présenter l'architecture globale de notre approche, à la fin on détaillera l'approche étape par étape.

I.1 Méta Modèle d'un Réseau Social

Notre travail consiste à analyser les données issues d'un réseau social, qui est constitué principalement d'utilisateurs et de leurs activités (publication, statut, post, partage...) et de relations entre utilisateurs (amitié, abonné...).

Tout d'abord, nous commençons par analyser les publications postées par des utilisateurs. La **figure I.1** représente le méta-modèle du réseau social « Twitter », pour but d'avoir un aperçu global des fonctionnalités du réseau étudié.

Un utilisateur peut publier un message, l'action de publication d'un message est dite « tweeter », le message publié est appelé « tweet », qui peut contenir des URLs, des hashtags et des mentions. Ce dernier peut être un tweet simple, un retweet ou une mention, dans le cas d'un retweet on dit que l'utilisateur a « Retweeter », alors que dans le cas d'une réponse on dit qu'il a « Mentionné ».

Un utilisateur peut signaler ou bloquer un autre utilisateur, comme il peut avoir des relations avec d'autres utilisateurs, il peut suivre et donc on dit qu'il a des « followings» et peut être suivi par d'autres utilisateurs, on dit qu'il a des « followers ».

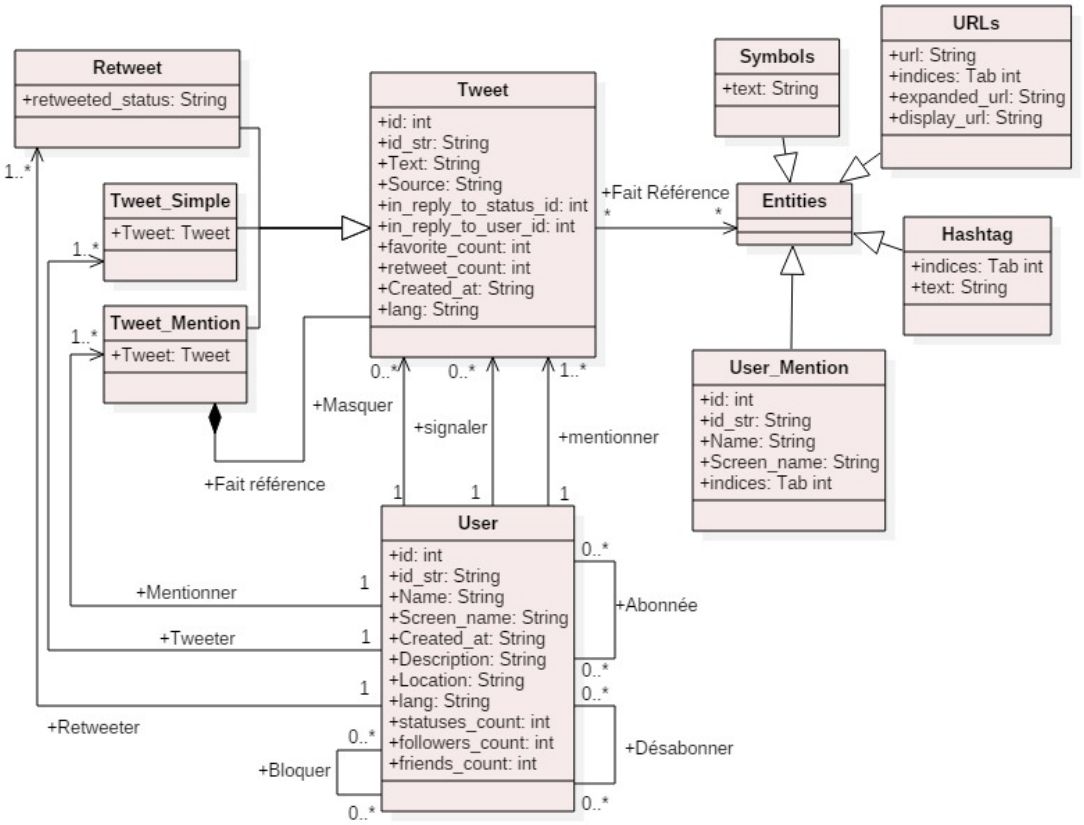

FIGURE I.1 – Méta-modèle du réseau social Twitter

I.2 Description de l'approche

Pour répondre à la problématique générale, nous avons conçu une approche qui permet de détecter les influenceurs sur les réseaux sociaux, en se basant sur le contenu des publications, et non pas sur les statistiques, car les travaux qui se reposent sur la popularité et l'impact de l'utilisateur dans le réseau ne tiennent pas compte réellement aux contenu des publications.
Notre contribution est composée de trois étapes essentielles :

- **Préparation du corpus :** Cette première étape permet de construire le corpus d'étude,en recherchant et récupérant des entités "publications" et ceci via un processus de recherche par mots-clés, en utilisant une ontologie qui aide à extraire les caractéristiques particulières de domaine.
- **Analyse de sentiment :** Cette étape permet d'analyser chaque publication, et leur attribuer un sentiment en utilisant des techniques d'apprentissage supervisé, pour but de l'exploité lors de la classification des posts.
- **Détection d'influenceurs :** Cette étape consiste à trouver les entités sources des opinions classés précédemment. Ces dernières sont évaluées selon leurs popularités et leurs impacts dans le réseau social. La **figure I.2** schématise notre approche.

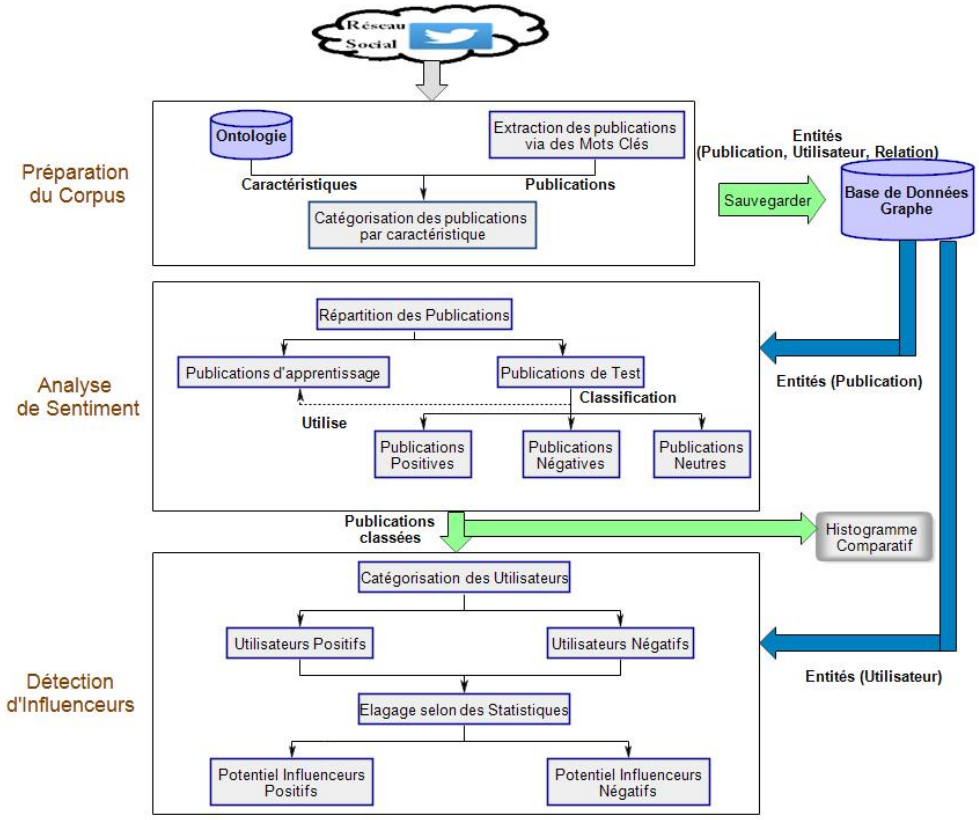

FIGURE I.2 – Schématisation de notre approche

Dans ce qui suit, nous allons détailler chaque étape de notre approche.

I.2.1 Étape 1 : Préparation du Corpus

Au niveau de cette étape, nous allons extraire les données pertinentes constituent notre corpus d'étude, qui est nécessaire pour l'analyse des sentiments. Ceci se fait en trois phases, l'extraction des entités « publications » du réseau social, à l'aide des caractéristiques qui seront extraites par une ontologie. Puis une catégorisation des publications par caractéristique aura lieu. A la fin une phase d'élagage est nécessaire pour garder que les publications pertinentes, car la récupération des entités engendre des entités superflues.La **figure I.3** résume cette étape de notre approche.

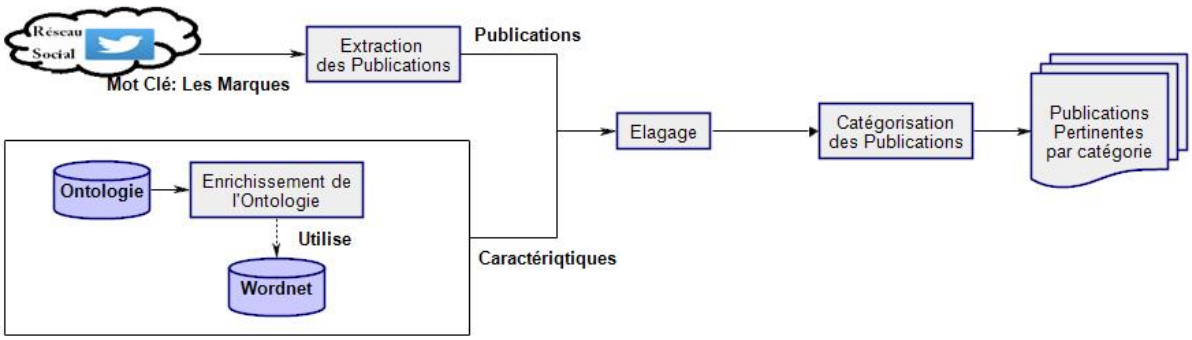

FIGURE I.3 – Étape de préparation du Corpus

I.2.1.1 Phase 1 :Extraction des entités « publications »

L'extraction des entités « publications » du réseau social se fait en temps réel. A l'aide de l'API du réseau social, on va recueillir et récolter des informations sur les publications, à partir de mot-clé « Marque ».

Cependant, afin de soustraire des entités plus raffiné et fournir une analyse des données plus élaborée, et aussi pour garantir de meilleurs résultats d'exécution de notre approche, nous allons prendre en considération les différentes caractéristiques du produit concerné.

Pour cela, On a opté pour une ontologie qui permet d'extraire les caractéristiques particulières d'un domaine, manipulable par les systèmes informatiques et interprétable par les êtres humains.

L'ontologie produite doit être complété par les synonymes et homonymes d'attributs détectés. En l'enrichissant la sémantique grâce à la base de données lexicale « WordNet » populaire. Par exemple, dans l'univers « smartphone » l'attribut « écran » peut aussi être exprimée comme « moniteur ».

La récupération des entités engendre des entités superflues, c'est pour cela que ces dernières subiront une phase d'élagage, pour garder que les publications pertinentes sur la base des concepts et propriétés incluses dans l'ontologie.

I.2.1.2 Phase 2 : Élagage des entités « publications »

Les données une fois récoltées subissent un traitement d'élagage, ce dernier consiste à :

- Éliminer les publications qui ne contiennent pas les caractéristiques définis dans l'ontologie.
- Éliminer les publications non conformes à la langue de l'utilisateur, dans notre cas, nous nous basons essentiellement sur les publications en anglais.
- Éliminer les publications dont le nombre de rediffusion ne dépasse pas un nombre N, car plus une publication est rediffusée plus elle est considérée comme pertinente.

I.2.1.3 Phase 3 :Catégorisation des entités « publications »

Après avoir filtré les publications non pertinentes, on passe à la phase de catégorisation des publications. Le but de la catégorisation automatique de textes est d'apprendre à une machine à classer un texte dans la bonne catégorie en se basant sur son contenu.
Habituellement, les catégories font référence aux sujets des textes, mais pour des applications particulières, elles peuvent prendre d'autres formes.

Dans notre cas, on regroupe automatiquement les données de chaque marque ensemble, afin d'avoir des catégories où chacune contient les publications transmettant une opinion sur l'une des caractéristiques d'une même marque.

Cette étape est cruciale, car elle constitue une étape de validation des données, et nous offre les éléments fondamentaux pour l'étude de notre approche, ceci dans le but de garantir une meilleure prestation.
Une fois le corpus d'étude bien défini, et avant de passer à l'étape d'analyse de sentiments, les entités « Publication, Utilisateur, Relation » pertinentes seront sauvegardés dans une base de données. Comme les réseaux sociaux sont généralement représentés sous forme de graphe, la sauvegarde sera dans une base de données orientée graphes (NoSQL). Il s'agit donc de stocker les données dans des nœuds et des arcs. Les publications et les utilisateurs sont représentées par des nœuds labellisés avec des informations sur la publication (son auteur, sa date de création, le nombre de retweet, ...etc), et sur l'utilisateur (son nom, sa localisation, ses followers, ses followings, le nombre de publication ...etc). Alors que les relations entre les publications et les utilisateurs sont représentés par des arcs. La **figure II.4** représente l'étape de sauvegarde des entités dans la base de données graphes.

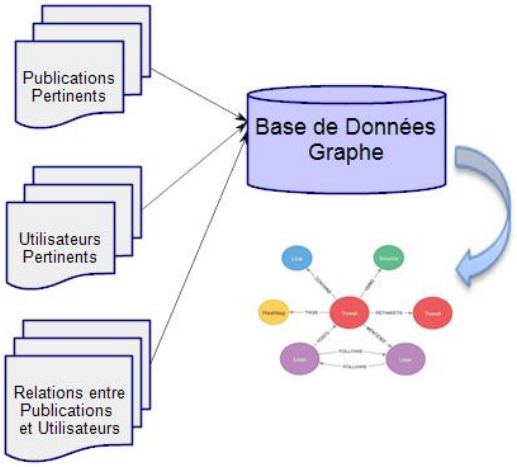

FIGURE I.4 – Sauvegarde des entités dans la base de données

I.2.2 Étape 2 : Analyse de sentiment

L'analyse de sentiment constitue un domaine de recherche populaire, en offrant des avantages à une variété de domaines. Dans notre cas cette étape consiste à regrouper les publications issues des réseaux sociaux selon l'émotion et le sentiment que dégagent les posts, que ça soit un sentiment positif, négatif ou neutre, et ainsi permettre aux entreprises de faire une étude comparative sur leur marque. Pour cela, on applique la méthode basée sur un corpus d'apprentissage afin d'entraîner un système automatique, et un corpus de test pour l'évaluation de performance du système automatique. La **figure II.5** présente un schéma explicatif de cette étape.

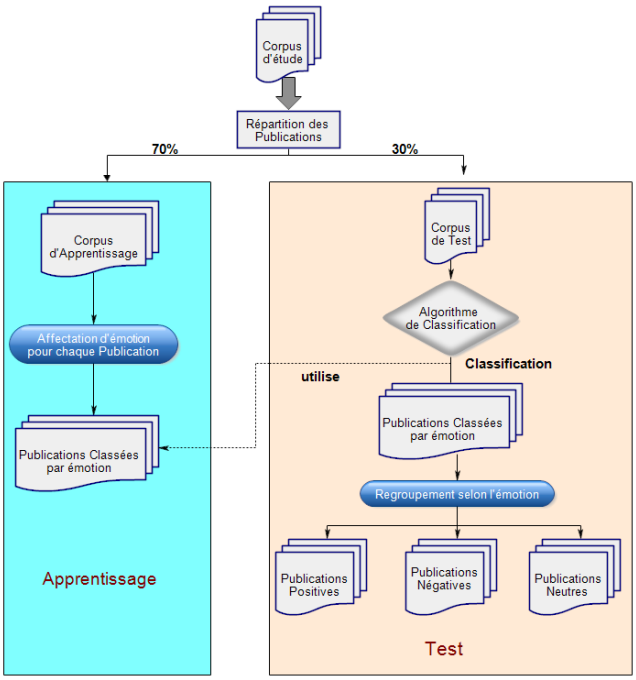

FIGURE I.5 – Étape d'analyse de sentiment

I.2.2.1 Phase 1 : Phase d'apprentissage

Cette première phase nécessite soit une intervention manuelle, soit l'utilisation d'un package « sentiment » de R, pour annoter les données d'apprentissage qui représente 70% des données récoltées de l'étape précédente.
Cette phase consiste à affecter l'émotion adéquate pour chaque publication, on prend en considération les émotions de base suivantes : la colère, le dégoût, la peur, la joie, la tristesse, la surprise dans le cas où la publication émerge une émotion, Or dans le cas où la publication ne donne aucun sentiment, on lui affecte « une émotion inconnu ». A la fin les publications annotées seront nécessaires pour la phase de test.

I.2.2.2 Phase 2 : Phase de test

Une fois les données d'apprentissage ont été classées selon les différentes émotions citées précédemment, et à partir d'un algorithme de classification, nous allons construire un classificateur appris à partir de ses données annoté. Ce classificateur va permettre de déterminer la classe de chaque publication du corpus de test, qui représente 30% du corpus d'étude.
Une fois les posts sont attribuer à chaque classe, nous passons a une étape de regroupement des publications, en rassemblent les données des quatre classes (colère, le dégoût, la peur, la tristesse) qui représentent un sentiment négatif, et les deux autres (la joie et la surprise) qui représentent un sentiment positif, dans deux classes distinctes de publications négatives et positives, quant aux publications qui n'expriment pas un sentiment précis seront regrouper dans une classe de publications neutre.
Après le classement des publications, un histogramme comparant les différentes marques peut être générer, qui indique les scores positifs, négatifs et neutres globaux de chaque caractéristique, dans le but d'améliorer la qualité de service.

I.2.3 Étape 3 : Détection d'influenceurs

À partir du moment que les publications ont été classé suite à une analyse de sentiments, nous estimons que ces dernières n'apporte pas préjudice aux entreprises, étant donné qu'on met abstraction à l'origine des critiques, c'est pour cela qu'on passe à une dernière étape de notre approche, qui consiste à repérer les potentiels utilisateurs susceptibles d'être source d'influence et centre de critique. Cette étape est composé de trois principales phases : Extraction des utilisateurs, puis une catégorisation des utilisateurs sera nécessaire, et finalement un élagage aura lieu afin d'éliminer les utilisateurs non influenceurs. La **figure II.6** montre les différentes phases de cette étape.

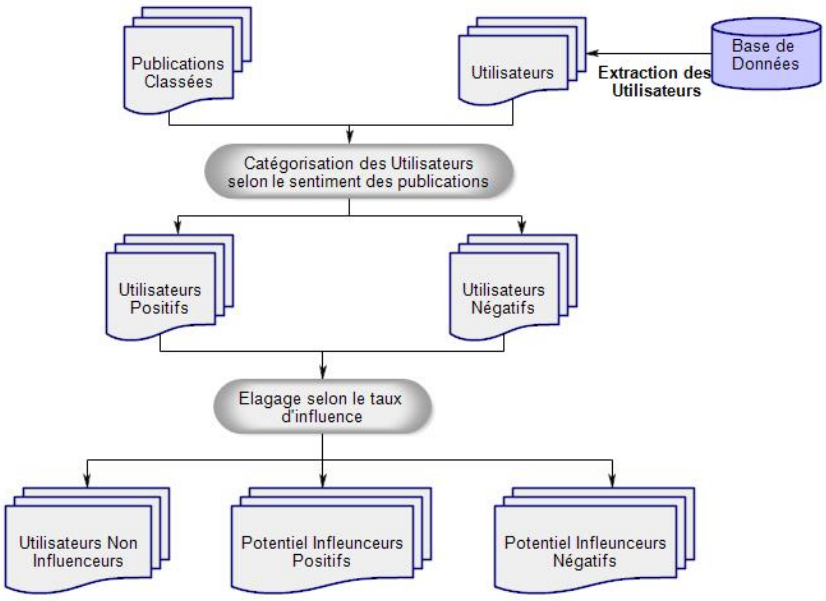

FIGURE I.6 – Étape de détection des influenceurs

I.2.3.1 Phase 1 : Extraction des utilisateurs

Une fois les entités « publication » sont classées comme présenté dans l'étape précédente, nous nous intéressants à la source qui a exprimé ce sentiment. Pour se faire, on va extraire de la base de données les utilisateurs approprié, afin de les analyser et détecter les potentiels influenceurs.

I.2.3.2 Phase 2 : Catégorisation des Utilisateurs selon le sentiment des publications

Cette phase consiste à classées les utilisateurs en trois principales catégories (Utilisateurs Positives, Négatives et Neutres), et cela en se basent sur le classement de leurs publications analysées dans l'étape précédente. Cependant les utilisateurs neutres n'apportent aucun préjudice dans notre travail d'où l'utilité de les ignoré.

I.2.3.3 Phase 3 : Élagage selon le taux d'influence

Après la phase de catégorisation des utilisateurs, ces derniers doivent subirent un filtrage pour garder que les utilisateurs susceptibles d'être influenceurs, et cela, en calculant un taux d'influence qui se base sur les deux critères suivants : la popularité et l'impact.

- **Attribution d'un score de popularité aux utilisateurs :** Dans cette partie, on doit attribuer pour chaque utilisateur un score de popularité, ce score dépend de la catégorie à laquelle appartient l'utilisateur et sa position dans la pyramide d'influence vu précédemment dans le chapitre 1 de la partie 1. Où chaque individu dont la position est plus proche du sommet se voit attribué une note plus importante que celui qui se trouve plus près de la base de la pyramide. L'évaluation de cette partie se fait suivant les étapes de **l'algorithme 1**.

```
Algorithme 1 : Calcul du Score de Popularité
Input  : Entités utilisateurs
Output : Entités utilisateurs avec Score popularité
for Utilisateur do
    if Utilisateur ∈ Catégorie1 then
        Score_Popularite ← Score_categorie1
    else
        if Utilisateur ∈ Catégorie2 then
            Score_Popularite ← Score_categorie2
        else
            Score_Popularite ← Score_categorie3
        end
    end
end
```

Algorithme 1 : Calcul Score Popularité

Cependant ce critère ne suffit pas pour dire qu'un utilisateur est un influenceur, d'où l'intérêt de voir les relations de cette utilisateur et son impact sur le réseau.

- **Calcul du score de l'impact des utilisateurs :** Dans cette partie, nous nous intéressons seulement aux relations d'un utilisateur dans le réseau social, car ses relations définissent « l'impact » que pourrait avoir un utilisateur sur un autre surtout qu'il existe une très forte corrélation entre le nombre d'abonnés et le d'abonnement représenté par un ratio qu'il semble être un indicateur très pertinent pour le repérage des influenceurs et donc une mesure d'influence importante pour notre étude.

  De ce fait, afin d'évaluer ce ratio (abonnés/ abonnement), nous procédons par le calcul des « arcs entrants » d'un utilisateur qui représente le nombre d'abonnement et des « arcs sortants » qui représente le nombre d'abonnés dans la base de données graphe.

  Nous avons regroupé les différents cas de figures des résultats de calcul du ratio dans le **Tableau I.1** suivant.

Valeur Approximative Du Ratio	Explication
<1	L'utilisateur a plus tendance à suivre les autres utilisateurs que d'être suivis.
>1	L'utilisateur est suivi par beaucoup de personnes.
=1	Il y un équilibre entre les suiveurs d'un utilisateur et ces suivis.

TABLE I.1 – Calcul du ratio (abonnés/ abonnement)

L'évaluation de cette partie se fait suivant les étapes de **l'algorithme 2**.

Algorithme 2 : Calcul du Score de l'impact
Input : Entités utilisateurs + Relation
Output : Entités utilisateurs avec Score impact
for *Utilisateur* **do**
  Ratio_Utilisateur= Nb_abonnés / Nb_abonnement ;
  Score_Impact ← Ratio_Utilisateur ;
end

Algorithme 2 : Calcul Score de l'impact des Utilisateurs

Une fois l'évaluation des deux mesures d'influence « popularité » et « impact » effectuées, le calcul du taux d'influence se fera comme le montre **l'algorithme 3** suivant.

Algorithme 3 : Calcul du Score d'influence
Input : Entités utilisateurs
Output : Entités utilisateurs avec Score d'influence
for *Utilisateur* **do**
  Score_Influence = Score_Popularité + Score_Impact ;
end

Algorithme 3 : Calcul Score d'influence

Suite à l'attribution de la valeur du score d'influence a chaque utilisateur, nous pouvons classer les utilisateurs comme suit :

- Si taux d'influence > N, alors on dit que c'est un potentiel influenceur.
- Si taux d'influence > N, alors on dit que ce n'est pas un influenceur.

A la fin de notre approche, nous aurons comme résultat un ensemble d'influenceurs positives et négatives d'une marque suivant ses caractéristiques, trié selon le taux d'influence attribué à chacun. Cependant, dans le cas où deux influenceurs se retrouvent avec le même taux d'influence, l'utilisateur pourra effectuer un deuxième tri, en prenant en considération la force d'influence, qui permet de classer l'influenceur diffuseur de tweet en première position avant l'influenceur rediffuseur.

I.3 Conclusion

Au niveau de ce chapitre, nous avons présenté notre approche pour la détection d'influenceurs dans les réseaux sociaux basée sur le contenu de leurs publications, qui se résume en trois parties. La première étape consiste en la préparation du corpus d'étude à partir de données issues du réseau social. Ensuite la deuxième étape consiste à l'étude du corpus préparé précédemment afin d'analyser le sentiment générer par la données récoltés suivent une classification supervisée. A la fin la détection d'influenceurs sera faite selon deux mesures d'influence : la popularité et l'impact de l'utilisateur dans le réseau.

Dans le prochain chapitre nous allons présenter la conception de l'outil que nous avons réalisé avec le langage UML.

CHAPITRE II

Modélisation de l'outil

Dans cette partie du mémoire nous allons présenter la conception de notre outil, qui permet de construire le corpus d'étude pour l'étape de l'analyse de sentiment, ainsi que la détection des influenceurs, qui se base sur la popularité de l'utilisateur et son impact dans le réseau social.
Les données seront stockées dans une base de données NoSQL dédiée aux graphes, nous avons choisi la base de données Neo4j connue pour son extensibilité et sa facilité d'utilisation.
La phase de conception permet de décrire de manière non ambiguë, le fonctionnement futur du système, afin d'en faciliter la réalisation, en utilisant un langage de modélisation. La conception de notre outil a été réalisée avec le langage UML (Unified Modeling Language), qui est un langage de modélisation graphique et textuel destiné à comprendre et à définir des besoins, spécifier et documenter des systèmes, concevoir des solutions et communiquer des points de vue.
Dans ce qui suit, nous présenterons les différents diagrammes de conception. Nous commençons par le diagramme de cas d'utilisation, qui est destiné à représenter les besoins des utilisateurs par rapport au système, suivit du diagramme de séquence pour décrire les scénarios de chaque cas d'utilisation et du diagramme d'activité. Et pour finir, nous présenterons notre diagramme de classe afin de présenter les classes et les relations entre elles.

II.1 Diagramme de cas d'utilisation

Nous allons enchainer notre conception par le diagramme de cas d'utilisation, qui décrit les fonctionnalités employées par les utilisateurs et permet de modéliser une interaction entre le système informatique à développer et un utilisateur, ainsi il permet de définir les besoins des utilisateurs et les limites du système. Il s'agit donc d'effectuer une analyse fonctionnelle de l'outil à réaliser.
Le diagramme de la **figure II.1**, illustre le cas d'utilisation générale de notre outil, qui englobe toutes les tâches que peut effectuer l'utilisateur.

Nous distinguons les cas d'utilisation suivants :

- Construction du corpus.
- Consultation des données récupérées.
- Lancement de l'apprentissage pour l'analyse de sentiment.
- Lancement de la classification pour l'analyse de sentiment.
- Consultation des résultats après l'analyse de sentiment.
- Détection des influenceurs.
- Génération du graphe des influenceurs.

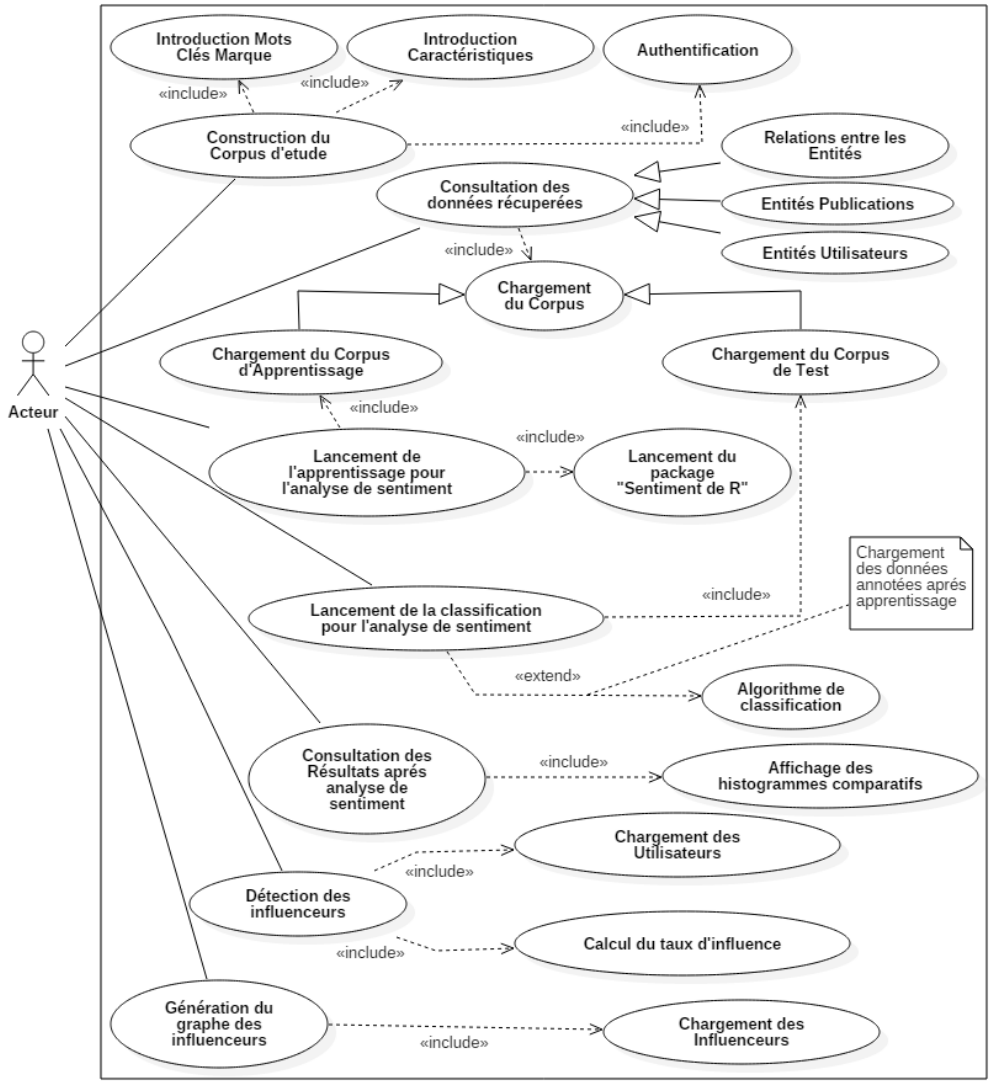

FIGURE II.1 – Diagramme de cas d'utilisation

Le **Tableau II.1** représente les cas d'utilisations avec leurs descriptions.

Cas d'utilisation	Description
Construction du corpus d'étude	Permet à l'utilisateur de construire le corpus d'étude à partir des données du réseau social. Pour cela, il doit d'abord s'authentifier (L'authentification d'un utilisateur nécessite son enregistrement au niveau du réseau Twitter, afin d'obtenir les clés d'accès pour l'extraction des données).Puis il doit introduire des mots clés, dans notre travail, nous avons opté pour quatre marque de smartphone (iPhone, Samsung Galaxy, HTC One, Nokia Lumia). De plus, l'utilisateur doit introduire les caractéristiques des produits, nous avons choisi les quatre caractéristiques du base des smartphones (Batterie, Ecran, Processeur, Caméra).
Consultation des Données récupérées	Le système donne la main à l'utilisateur pour la consultation des entités extraites du réseau social. Ces entités sont des publications, des utilisateurs et des relations entre eux.
Lancement de l'apprentissage pour l'analyse de sentiment	Permet à l'utilisateur de lancer le processus d'apprentissage par le biais du corpus d'apprentissage, et le lancement du package « sentiment de R » qui affecte une émotion pour chaque publication.
Lancement de la classification pour l'analyse de sentiment	Au niveau de cette étape, l'utilisateur peut lancer la classification des publications, qui sera réalisée en chargeant le corpus de test, et le lancement d'un algorithme de classification.
Consultation des résultats après l'analyse de sentiment	Permet à l'utilisateur de consulter les résultats, en affichant les différents histogrammes comparatifs.
Détection des influenceurs	Dans cette étape, l'utilisateur aura la main pour détecter les influenceurs,par le calcul du taux d'influence et cela, après avoir chargé les utilisateurs pertinents.
Génération du graphe des influenceurs	Permet à l'utilisateur de générer le graphe des influenceurs, en chargeant les potentiels influenceurs à partie de la base de données.

TABLE II.1 – Description des cas d'utilisation

II.2 Diagramme de séquence

Le diagramme de séquence permet de décrire les scénarios de chaque cas d'utilisation, en mettant l'accent sur la chronologie des opérations et des interactions entre acteur et objets du système manipulés par l'utilisateur. Nous avons choisi de décrire les cas d'utilisation les plus nécessaires. Les diagrammes de séquence de notre outil sont représentés dans ce qui suit.

II.2.1 Diagramme de séquence d'authentification

Le diagramme de la **figure II.2** décrit le scénario possible lors de l'identification de l'utilisateur, qui doit être inscrit au niveau du réseau social Twitter, afin d'obtenir des clés d'accès pour l'extraction des entités. La phase d'authentification sera détaillée dans le prochain chapitre.

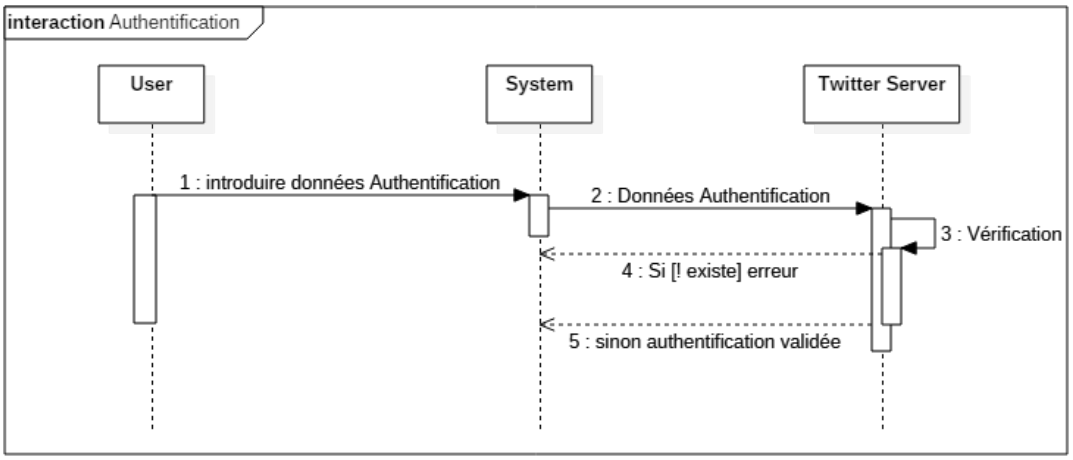

FIGURE II.2 – Diagramme de séquence d'authentification

II.2.2 Diagramme de séquence de préparation du corpus d'étude

Lors de cette étape, et après la phase d'authentification, l'acteur doit extraire les entités à partir du réseau social, en introduisant les mots clés identifiant les marques, ainsi les différentes caractéristiques. Une fois les entités récupérées du réseau social, une opération d'élagage des entités récupérées est effectuée afin d'éliminer les publications non conforme à la langue, dans notre travail on s'intéresse au publications en anglais seulement, et aussi pour éliminer les posts qui ne contiennent pas les caractéristiques définis, puis une catégorisation des entités par marque aura lieu avant d'entamer le stockage dans la base de données.
La **figure II.3** illustre l'opération de construction du corpus.

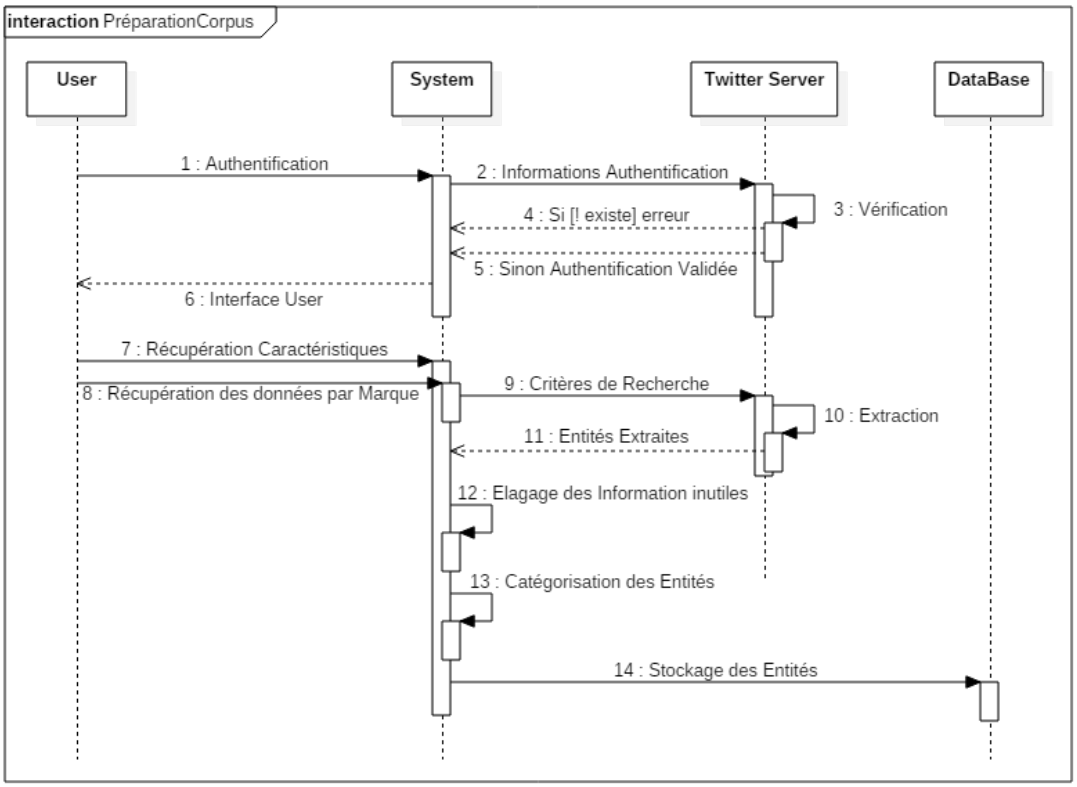

FIGURE II.3 – Diagramme de séquence de préparation du corpus

II.2.3 Diagramme de séquence de consultation des entités

Après avoir préparé le corpus d'étude et sauvegarder les entités dans la base de données, cette fonctionnalité permet à l'acteur de consulter les différentes entités extraites. Il peut donc consulter les différentes publications extraites, les utilisateurs et les relations entre le tweet et son auteur. La **figure II.4** représente cette étape.

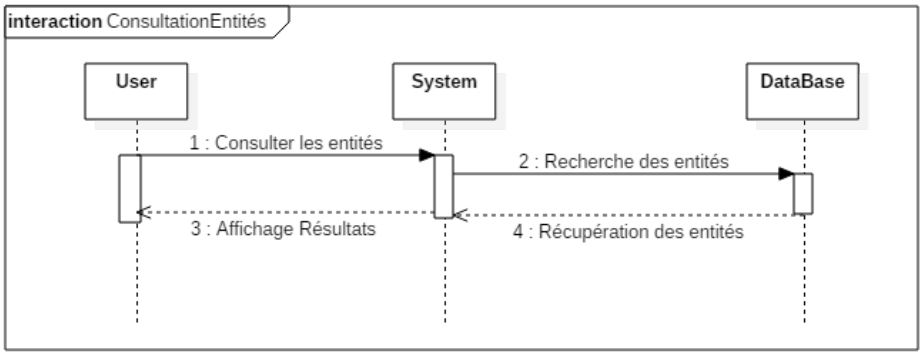

FIGURE II.4 – Diagramme de séquence de consultation des entités

II.2.4 Diagramme de séquence de l'analyse de sentiment

Cette opération permet à l'acteur d'analyser le sentiment des différentes publications, et cela en passant d'abord par lancer un apprentissage, en utilisant le package « sentiment de R » pour annoter automatiquement les entités par l'émotion adéquate, et cela après avoir fixé un pourcentage pour les données d'apprentissage, ensuite le reste du corpus d'étude sera consacré à la phase de classification, en utilisant les données déjà annoté et un algorithme de classification, dans notre travail on a opté pour le classifieur bayésien qui est très utile pour la classification des textes. La **figure II.5** montre les différentes interactions de cette étape.

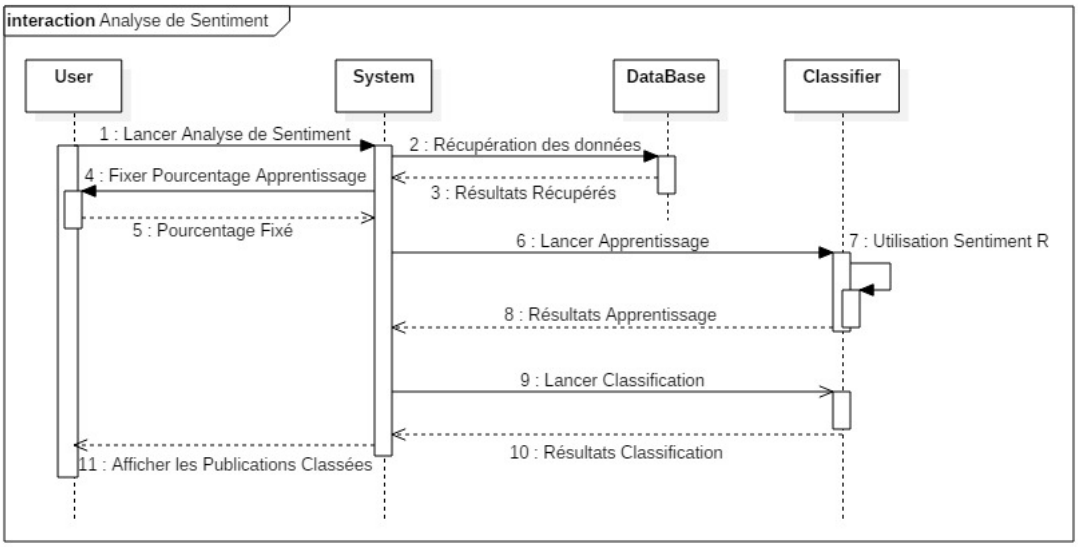

FIGURE II.5 – Diagramme de séquence de l'analyse de sentiment

II.2.5 Diagramme de séquence de détection des influenceurs

Cette fonctionnalité permet à l'acteur de détecter les utilisateurs influenceurs, en calculant un taux d'influence et cela à base des utilisateurs source des publications classée précédemment et de deux critères d'influence :la popularité de l'utilisateur et son impact dans le réseau.

- Dans le cas où le taux d'influence est supérieure à N, alors cette utilisateur est désigné comme un potentiel influenceur.
- Dans le cas où le taux d'influence est inférieure à N, alors cet utilisateur est désigné comme non influenceur.

La **figure II.6** illustre cette fonctionnalité.

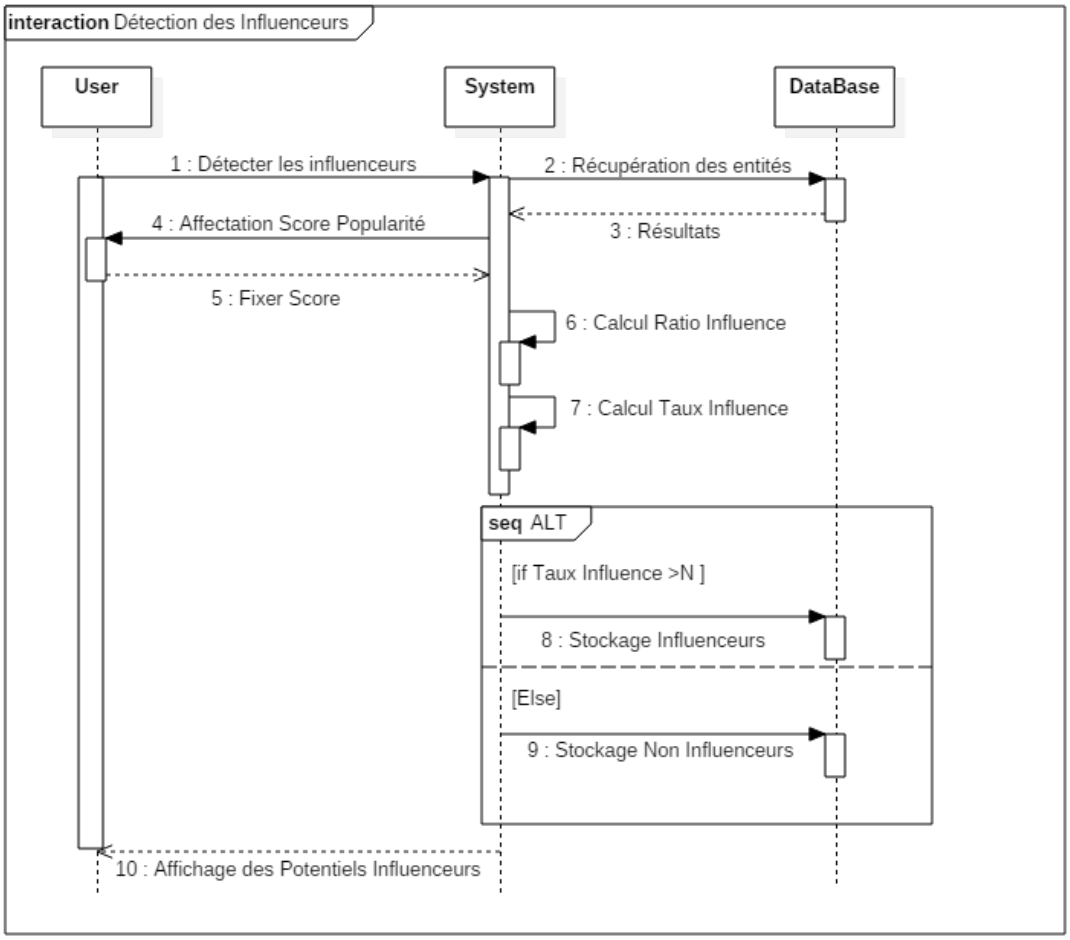

FIGURE II.6 – Diagramme de séquence de détection des influenceurs

II.3 Diagrammes d'activité

Le diagramme d'activité est destiné à représenter le comportement interne d'une méthode ou d'un cas d'utilisation. Ce diagramme donne une vision globale des enchaînements des activités propres à une opération ou l'exécution d'un mécanisme.

Dans notre cas, on va illustrer le fonctionnement du processus de manière générale à savoir les trois étapes : préparation du corpus d'étude, l'analyse de sentiment et la détection des influenceurs. La **figure II.7** illustre le déroulement d'étapes séquentielles de notre outil.

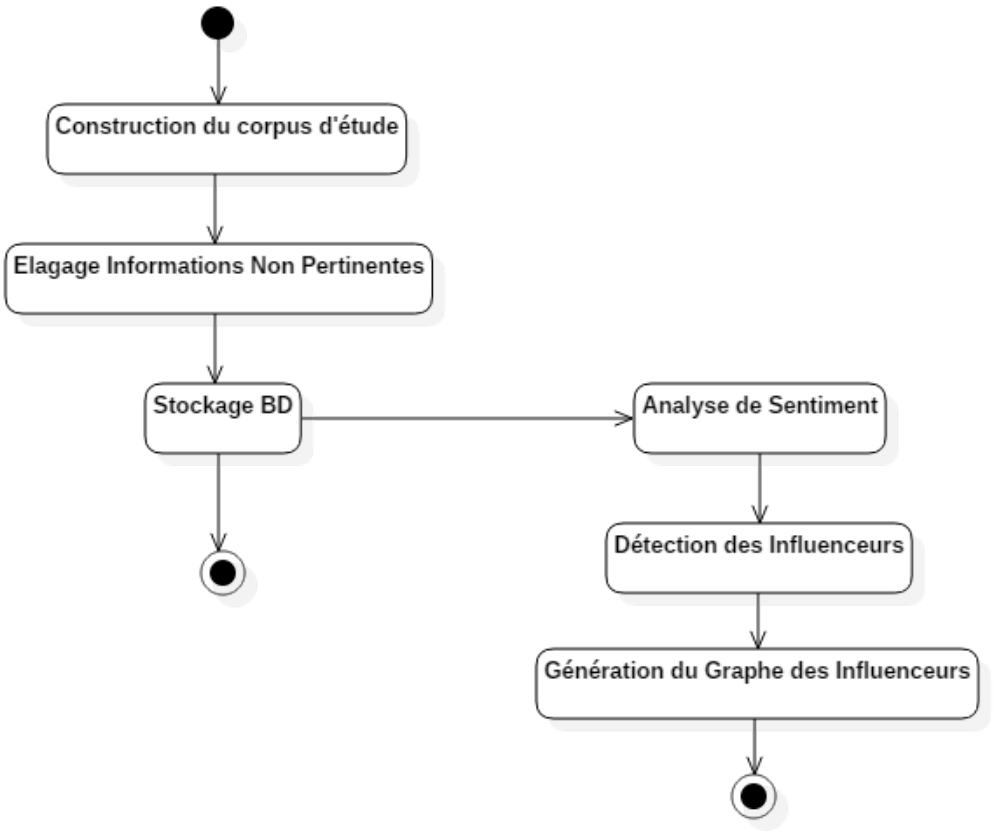

FIGURE II.7 – Diagramme d'activité du système

II.4 Diagramme de classe

Un diagramme de classe est une collection d'éléments qui décrit de manière générale la structure d'un système, à savoir la structure interne des éléments et leurs relations les uns par rapport aux autres.
Il représente la description statique du système en intégrant dans chaque classe la partie dédiée aux données et celle consacrée aux traitements.
La **figure II.8** représente le diagramme de classe de notre outil.

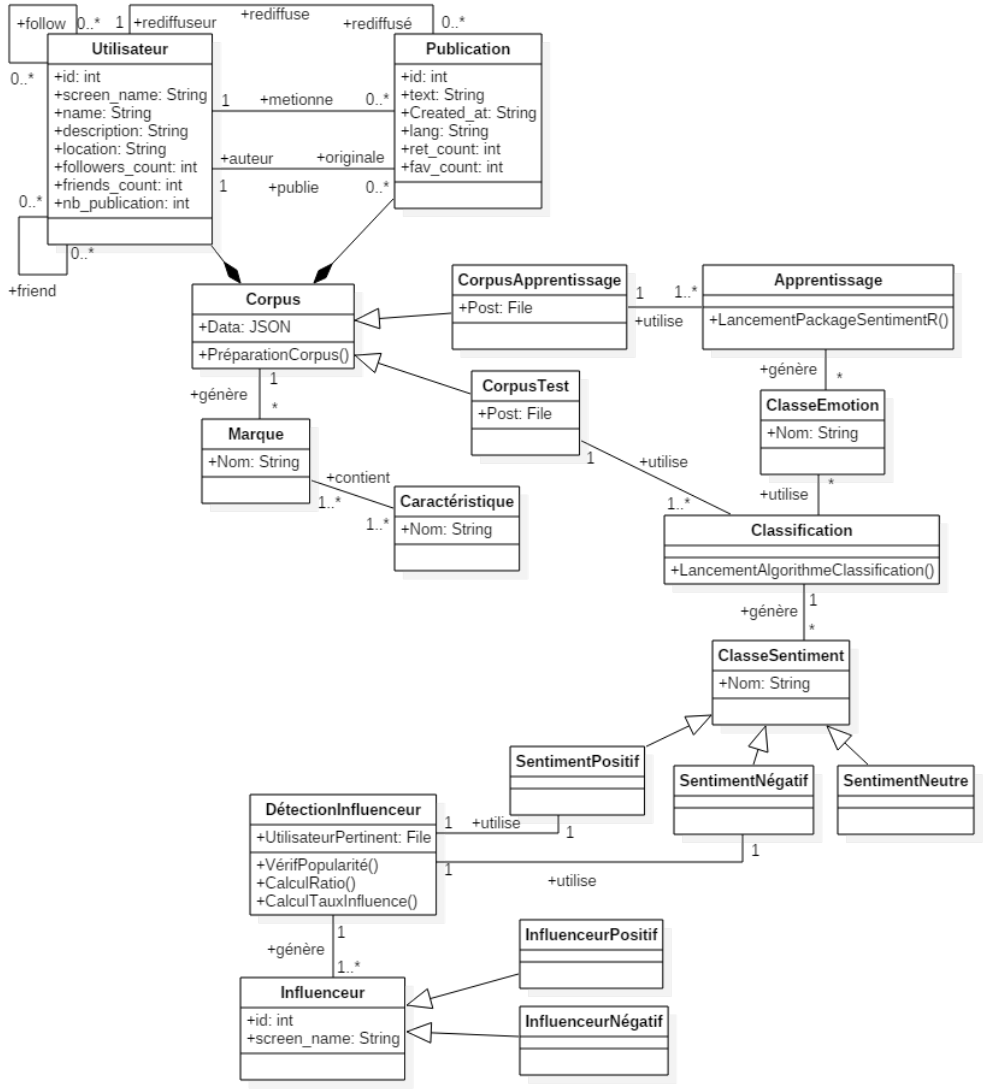

FIGURE II.8 – Diagramme de classe

II.5 Conclusion

Ce chapitre a été consacré à la modélisation de notre approche réalisée en UML. Après avoir effectué une étude préliminaire, passant par les étapes du processus de développement en faisant appel aux diagrammes d'UML, alors dans le chapitre suivant nous allons passer à la dernière étape de notre projet qui est l'implémentation d'un outil pour la détection des influenceurs.

CHAPITRE III

Implémentation de l'outil

Après avoir présenté notre approche de détection des influenceurs dans le premier chapitre, nous allons à présent procéder à la description détaillée de notre implémentation au cours de ce chapitre.
Dans un premier lieu, nous allons décrire l'environnement du travail et les principales technologies utilisées pour le développement de notre outil. Dans un second lieu, nous présenterons le travail réalisé, qui est organisé en trois principaux modules :

- Un module de préparation de corpus, ce module comprend l'accès à l'API twitter et la récupération des données à partir de Twitter par le biais de mots-clés, fournis par l'utilisateur à l'aide des caractéristiques extraites d'une ontologie.
  Ensuite la sauvegarde et la représentation des entités récoltées se fait dans une base de données, notre choix s'est posé sur une base de données orientée graphe Neo4j.
- Un module d'analyse de sentiment, ce module prend en entrée les données récoltées précédemment, puis retourne en sortie des tweets classés selon leur sentiment, en se basant sur un apprentissage automatique.
- Un module de détection d'influenceurs, qui permet de calculé le taux d'influence a base des deux critères d'influence : la popularité et l'impact de l'utilisateur dans le réseaux sociale.

III.1 Twitter API

Les utilisateurs de Twitter génèrent des millions de tweets par jour, certains de ces tweets sont disponibles à travers des API (Application Programming Interface) publiques, destinée à l'attention des développeurs qui désirent proposer aux consommateurs de nouvelles manières d'utiliser Twitter.
Pour chaque « tweet » posté sur Twitter, l'API fournit à ses utilisateurs tout un ensemble d'informations dont, entre autres, la date et l'heure du tweet, son contenu, le fait qu'il soit en réponse d'un autre tweet, l'identification de l'expéditeur du tweet, la date de création de son compte, le nombre de suiveurs, de suivis et de tweets qu'il comptabilise depuis la création de son compte. L'API permet également d'obtenir des renseignements sur les relations (suiveur, suivi) qu'entretiennent les utilisateurs de Twitter entre eux, et fournit d'autres champs qui permettent de créer

des programmes pour suivre automatiquement plusieurs personnes en même temps, arrêter de les suivre, les bloquer. . . etc.
Dans ce qui suit, nous présenterons les différentes API utilisées, ainsi, les conditions d'accès à ses derniers.

III.1.1 Les API de Twitter

Dans le cadre de notre travail, nous avons principalement utilisé l'API de Twitter dans le but de récupérer des tweets et leurs utilisateurs. Les APIs d'accès aux données de Twitter peuvent être classées selon deux types :

III.1.1.1 REST API

Cette API fournit les données à un format XML (Extensible MarkupLanguage) qui permet deobtenir des données balisées, c'est-à-dire pré formatées dans des champs arborescents, afin de faciliter l'échange des données entre Twitter et les utilisateurs de l'API.[J.P and A, 2009]
Ce point d'accès permet d'accéder aux tweets historisés par Twitter. Il est basé sur l'architecture « REST » couramment utilisée pour la conception des APIs Web. Ces APIs utilisent la stratégie « PULL » pour la récupération des données. Pour recueillir des informations à partir de Twitter, un utilisateur doit explicitement envoyer une demande (requête).
Cependant, Twitter impose des limites à l'utilisation de son API REST. Ces limites définissent le nombre de tweets envoyés par jour (1000), le nombre de courriers privés envoyés aux autre utilisateurs (250 « direct messages » / jour), et enfin – le plus important en ce qui nous concerne – le nombre maximum de requêtes dans l'API (150 / heure) avec une limite de 800 tweets collectés par requête (soit un maximum de 150*800 = 120 000 / heure). [J.P and A, 2009]

III.1.1.2 Stream API

Cette API fournit des données au format JSON (Java Script Object Notation). L'utilité de cette API est de permettre une sélection des contenus présents sur twitter sur n'importe quel mot clé. Aussi, dans l'exemple où le chercheur voudrait analyser uniquement les échanges contenant un (ou plusieurs) mot(s) précis (nom de marque par exemple), il ne serait pas obligé de collecter tous les tweets publiés pour ensuite sélectionner uniquement ceux contenant le(s) mot(s). Il collecte d'emblée uniquement les tweets qui correspondent à la requête. Cette API permet en outre de sélectionner les tweets en fonction de langage déclaré par l'utilisateur.[J.P and A, 2009]
Elles fournissent un flux continu d'informations publiques de Twitter, et est donc la livraison est en temps réel. Ces APIs utilisent la stratégie « PUSH » pour la récupération de données (en fournissant les mises à jour sans intervention de l'utilisateur).
Dans notre cas nous avons opté pour l'utilisation de l'API Streaming de twitter car comme notre approche se base sur la récolte de tweets en temps réels, il est préférable de suivre les tweets comme ils se produisent, de plus, notre but est de récolter des données par marque. Ce qui rend l'API Streaming d'autant plus approprié pour notre cas.

Le point commun entre les deux types est qu'ils fournissent l'autorisation seulement par l'authentification OAuth.

III.1.2 Accès à Twitter API

Avant de tenter de récupérer les tweets, une application doit être créée avec le compte de l'utilisateur, et enregistrée sur Twitter developpers,cette application permet aux utilisateurs de rechercher les tweets. Une fois l'application est créée, Twitter fournit deux clés d'accès à l'application, les clés d'accès secrète « Consumer Key » et « Consumer Secret ».
Twitter vérifie l'identité de l'utilisateur et lui délivre un PIN. L'utilisateur fournit ce PIN à son application pour demander un « Access Token » et un « Access Secret » unique à cet utilisateur. En utilisant les jetons d'accès « Access Token » et « Access Secret » l'application authentifiée l'utilisateur auprès de Twitter, et lance des appels API au nom de l'utilisateur.
Pour permettre aux applications de fournir ces informations, l'API de Twitter repose sur le protocole OAuth.
OAuth est un standard libre qui permet d'autoriser une application client à utiliser l'API sécurisée d'une autre application pour le compte d'un utilisateur.L'intérêt majeur d'OAuth vient du fait que l'utilisateur n'a plus besoin de fournir ses informations d'identification à une application tierce car la connexion se passe sur l'application de l'API. Cela suppose que l'utilisateur lui a à priori fait confiance.Le processus d'authentification à Twitter est illustré dans la **figure III.1**.

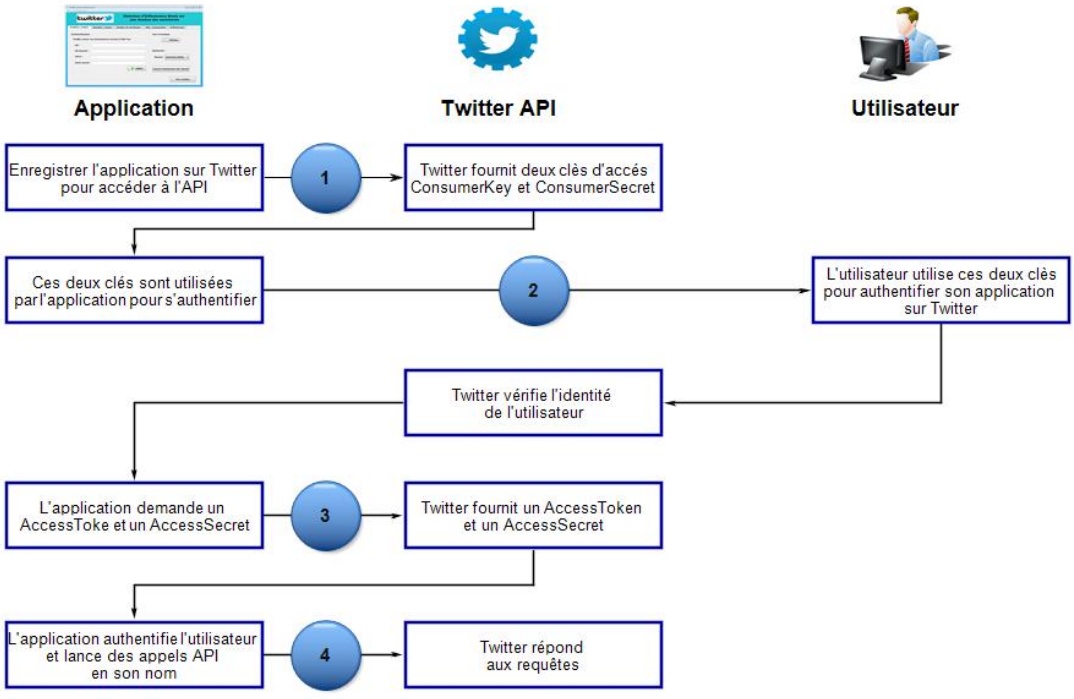

FIGURE III.1 – Les étapes d'authentification sur Twitter API

III.2 Outils de développement

Dans cette section, nous présentons les différentes plateformes logicielles nécessaires pour le fonctionnement de notre système. Nous passerons en revue les environnements de développement utilisés ainsi que les arguments relatifs aux choix de leurs utilisations. A la fin nous présenterons la base de données orientée graphe Neo4j.

III.2.1 Langages de programmation

Nous avons choisi d'implémenter notre application avec le langage multiplateforme « Python », et le langage R qui est un logiciel libre de traitement des données et d'analyse statistiques.

III.2.1.1 Python

Pour assurer un meilleur déploiement de notre système de recherche d'influenceurs dans un réseau social, nous avons décidé de le rendre indépendant vis-à-vis des différentes plateformes (systèmes d'exploitation) qui opèrent sur la machine. L'implémentation de notre application doit prendre en considération un paramètre primordial : la portabilité. Pour cela, nous avons fait le choix d'un langage multiplateforme qui est : Python et plus précisément la version 2.7.
Python est un langage portable, dynamique, extensible, gratuit, qui permet une approche modulaire et orientée objet de la programmation. Nous avons opté pour ce langage grâce à sa facilité d'utilisation et de ses fonctions dans le domaine de traitement de texte.

Les modules Python utilisés
Python est un langage de programmation interprété, simple et puissant, il permet d'écrire des scripts très simples mais grâce à ses nombreuses bibliothèques, il aide le développeur à travailler sur des projets plus ambitieux. Dans le **tableau III.1** suivant nous expliquons les modules essentiels qu'on a utilisés.

Module	Explication
Tweepy	C'est un module qui rend l'utilisation de twitter de streaming api plus facile, en manipulant l'authentification, la connexion, la création et la destruction de la session, la lecture des messages entrants, et partiellement le routage des messages. Il soutient l'accès à Twitter via l'authentification de base avec la méthode OAuth.
Py2neo	Le package py2neo contient des classes et fonctions requises pour interagir avec un serveur Neo4j. Le plus important d'entre eux est la classe graphique qui représente une instance de base de données graphique Neo4j et d'effectuer des requêtes en langage cypher grâce à son module cypher.
TextBlob	La bibliothèque de traitement de données textuelles. Il fournit une API simple pour la plongée dans le traitement commun du langage naturel (NLP) des tâches telles que l'extraction d'expression, l'analyse des sentiments, la classification, la traduction. Dans notre cas, on a utilisé ce module afin d'avoir le classifieur bayésiens.

TABLE III.1 – Les modules Python utilisés

III.2.1.2 Le langage R

R est un langage de programmation interactif interprété et orienté objet contenant une très large collection de méthodes statistiques et des facilités graphiques importantes.
Un des avantages les plus importants de R est qu'il fournit à travers plusieurs packages des fonctions, tel que :

- la mise en œuvre d'une classification à l'aide d'un bayésien naïf.
- la manipulation des données.
- Les logiciels graphiques sont plus puissants pour l'analyse visuelle de données.

Le package du sentiment de R est utilisé pour classer la polarité des tweets et il utilise un ensemble de données intégré pour classer approximativement les émotions en six catégories comme la colère, la joie, le dégoût, la peur, la tristesse et la surprise, ce package est utilisé aussi pour faire toutes les techniques de prétraitement comme l'enlèvement des numéros, la ponctuation, les mots vides ... etc.

III.2.2 Environnements de développement

Dans ce qui suit, nous présentons les différents environnements de développement.

III.2.2.1 Éditeur d'interface graphique : PyQt

PyQt est l'une des options de Python libre pour la programmation de l'interface graphique, édité par la société Riverbank Computing qui propose à la fois une version sous licence GPL et une licence commerciale Il permet de lier le langage Python avec la bibliothèque Qt et créer des interfaces graphiques.
Une extension de QtDesigner (utilitaire graphique de création d'interfaces Qt) permet de générer le code Python d'interfaces graphiques.

III.2.2.2 Éditeur d'ontologie : Protégé

Protégé est un éditeur d'ontologies distribué en open source par l'université en informatique médicale de Stanford. Il est très populaire dans le domaine du Web sémantique et au niveau de la recherche en informatique. C'est un outil qui permet de gérer des contenus multimédias, interroger, évaluer et fusionner des ontologies, et peut lire et sauvegarder des ontologies dans la plupart des formats d'ontologies : RDF, RDFS, OWL,... etc.
Protégé n'est pas un outil spécialement dédié à OWL, mais un éditeur hautement extensible, capable de manipuler des formats très divers.

III.2.2.3 Environnement de travail

Nous avons travaillé sur le système d'exploitation Windows 7, 64 bit, installé sur des machines doté d'un processeur Intel i5, CPU 2.50 Ghz, et une mémoire vive de 8 GO.

III.3 Base de données orientée graphes Neo4j

Neo4j est un système de gestion de bases de données orienté graphe développé en Java par la société suédo-américaine NeoTechnology. À la différence des systèmes classiques, son approche n'est pas fondée sur l'algèbre relationnelle mais sur la théorie des graphes. Les données sont stockées de manière assez libre (sans modèle prédéterminé) dans des nœuds reliés entre eux par des relations (des arcs porteurs d'une sémantique forte).
Neo4j utilise un langage d'interrogation spécifique CYPHER, conçu pour être assez intuitif. Cependant, Neoj4 se compose principalement des trois concepts présenté dans le **tableau III.2** suivant :

Concept	Description
Nœuds	Les nœuds sont des enregistrements composés de propriétés de type clé/valeur. Généralement, ils représentent une entité du modèle, ils peuvent contenir des propriétés et peuvent avoir 0 à plusieurs labels.
Relations	Les relations représentent les liaisons entre les nœuds du graphe, elles possèdent nécessairement un nœud de départ et un nœud d'arrivée. Les relations sont toujours orientée (entrante ou sortante), cependant, elles peuvent être traversées dans les deux directions. Un nœud peut être relié à lui-même. Toute relation à un type qui peut être vu comme un label.
Propriétés	Les nœuds et les relations peuvent avoir des propriétés qui sont des couples clé-valeur. La clé est toujours de type String, la valeur peut être de type primitif (boolean, byte, int… etc) ou un tableau de primitifs. NULL n'est pas une valeur valide, il est modélisé par l'absence de clé.

TABLE III.2 – Les concepts du Neo4j

Les bases de données Neo4j, et d'une manière plus générale les bases de données orientées graphes, sont particulièrement adaptées dans des contextes où les données sont fortement connectées et organisées selon des modèles complexes. La structure de l'entité (nœud ou relation) y est définie au moment du stockage de la donnée (et non préalablement comme les tables d'une base de données relationnelle), ce qui lui confère une très grande flexibilité. En outre, elles présentent des performances exceptionnelles en termes de lecture et de parcours de données dans le graphe.[ref, f]

III.3.0.4 Avantages de Neo4j

- Obtention des performances exceptionnelles pour la recherche de chemin : plus court chemin, détection de boucles, identification de sous-graphes, …. ou de manière plus concrète : calculs de trajets, mais aussi détection de fraude, recommandation, réseaux sociaux, etc…
- Des livrables en temps record : La Modélisation d'une base de donnée est simple et facile. Les entreprises peuvent capturer rapidement toutes sortes de données, structurées, semi-structurées et déstructurées et ainsi les stocker dans Neo4j. Ceci résultant dans une réduction des temps de développement, une réduction de coûts de maintenance.
- Manipulation des données fortement connectées.
- Supporte un modèle complexe et flexible, qui permet l'exécution de requêtes complexes avec une haute performance.
- Analyse avec une grande profondeur les relations entre les données.

- Fiabilité garantie par le respect et la compatibilité avec les propriétés ACID.
- Extensible (peut contenir des milliards de nœuds, relations et propriétés).
- Simple et accessible par une interface REST ou une API Java orientée objet.
- Haut disponibilité via la sauvegarde en ligne.
- Facilité de changement du schéma et des migrations de données sans contraintes.
- Neo4j peut être utilisé à la fois comme base embarquée sans aucun travail supplémentaire d'administration et comme un serveur à part entière.
- Chargement des données ultra rapide en intégrant le chargement de données dans le langage Cypher.

La base de données Neo4j et comme toute base de données utilise un langage pour manipuler et échanger des données, c'est là qu'intervient Cypher.

III.3.0.5 Langage d'interrogation Cypher

Neo4j a développé Cypher, qui est un langage de requêtes déclaratif orienté graphes proche du langage SQL, optimisé pour la lisibilité et la facilité d'expression humaine pour les développeurs, administrateurs et les experts de domaine. Il est également très intuitif, basé sur la représentation graphique du modèle graphe.

Pour expliquer clairement l'organisation d'une base de données sur Neo4j et l'utilisation de Cypher, nous allons l'expliquer par quelques exemples pour la création des utilisateurs et des tweets dans le réseau social Twitter, ainsi les relations entre eux :

- Pour créer une base de données, il faut commencer par créer les objets, dans notre cas c'est les utilisateurs et les tweets. Pour se faire, une syntaxe est mise en place :

```
CREATE { id :1 ,screen_ name :'Neo' } ) ;
CREATE (t : Tweet {id : 21, text :'Bienvenue dans notre application ! } ) ;
```

  – « **CREATE** » : permet d'indiquer qu'on va créer un nœud.
  – « **u** » : est un alias pour le nœud.
  – « **User** » : représente l'intitulé du nœud.
  – « **id** » et « **screen_name** » : représentent les propriétés du User.

- Pour créer la relation « utilisateur-tweet », la syntaxe suivante est mise en place :

```
MATCH (u :User {screen_name :'Neo'}), (t : Tweet {id : 21, text :' Bienvenue
dans notre application !})
CREATE (t) – [TWEETED BY] -> (u)
```

  – « **MATCH** » permet de collecter les données des nœuds pour lesquels la relation doit être créé.
  – « **CREATE** » permet de créer la relation.

III.4 Architecture global de notre outil

Dans ce qui suit, nous détaillerons les différentes composantes de notre outil en présentant quelques interfaces graphiques relatives aux fonctionnalités offertes. La **figure III.2** illustre l'architecture globale de notre application.

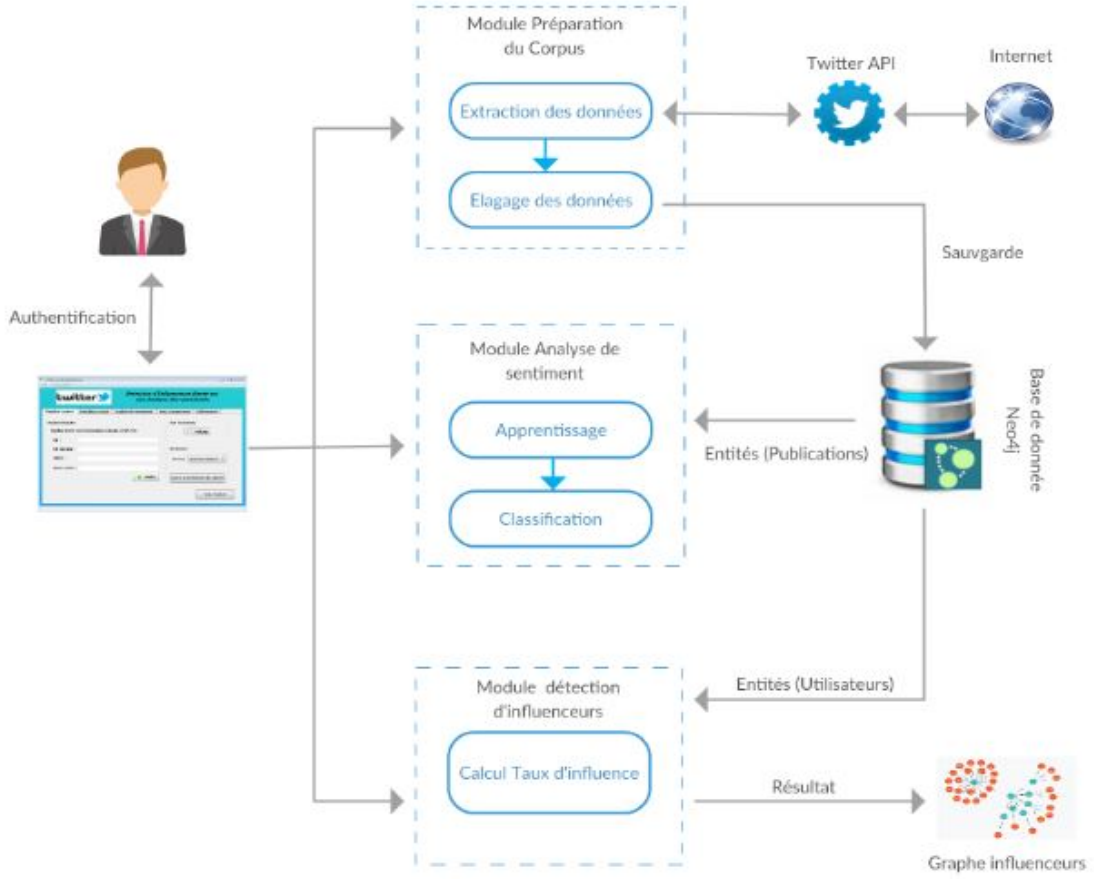

FIGURE III.2 – Architecture globale de notre application

Pour plus de flexibilité, nous avons adopté une architecture modulaire pour l'implémentation des phases décrites dans le chapitre 1 de la partie 2. Nous avons implémenté plusieurs modules, comme illustré dans la **figure III.2**. Le système implémenté se compose donc des trois modules suivants :

III.4.1 Module Préparation du corpus

Lors de ce module, l'utilisateur doit préparer les données pour le traitement, cette étape est importante, car elle constitue une étape de validation des données qui seront apte à être analysées et étudiées dans la suite de notre approche, et ceci dans le but de garantir de meilleurs résultats.Ce module est composé de deux sous-modules :

III.4.1.1 Sous module d'extraction des données

En premier lieu, l'utilisateur doit créer une ontologie afin d'extraire les caractéristiques du domaine. Diverses méthodes et de nombreux outils sont disponibles pour le développement de l'ontologie. Dans notre cas, les objets, les attributs et les relations entre eux sont identifiés pour le domaine des smartphones. Dans le présent document, nous utilisons l'un des outils les plus populaires appelés protégé 4.3 pour la création de l'ontologie. La **figure III.3** montre l'ontologie créée.

Une fois l'ontologie créée, l'utilisateur doit s'authentifier pour procéder à la recherche et à la récupération des informations à partir du réseau Twitter, ensuite l'extraction des données se fera suivant les mots clés représentant les marques définie dans l'ontologie.La **figure III.4** montre cette opération.

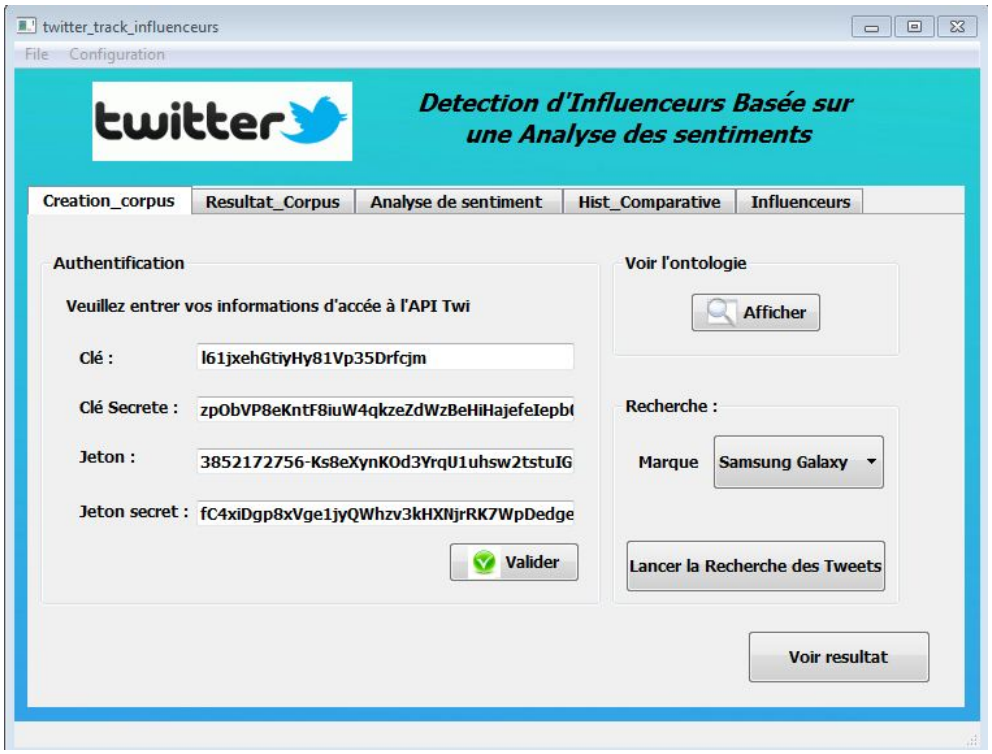

FIGURE III.3 – Interface du module de préparation du corpus

III.4.1.2 Sous module d'élagage des données

Après avoir collecté les données nécessaires, on passe par un filtrage afin de garder que des entités pertinentes pour l'étape suivante. L'élagage consiste à éliminer les données qui ne contiennent pas les caractéristiques spécifiées dans l'ontologie, ainsi éliminer les entités non conforme à la langue anglaise.

Notre corpus contient 1487 Tweets, représentant les quatre marques suivantes : iPhone, Samsung Galaxy, Nokia Lumia et HTC One,repartie comme nous le montre le **tableau III.3** suivant :

Marque	Nombre_Tweet
Samsung Galaxy	383
Appel iPhone	347
Nokia Lumia	295
HTC One	463

TABLE III.3 – Corpus d'expérimentation

Une fois les données récolter l'utilisateur peut visualiser le corpus générée et peut dissocier les publications de leurs utilisateurs. La **figure III.5** Suivante illustre cette phase.

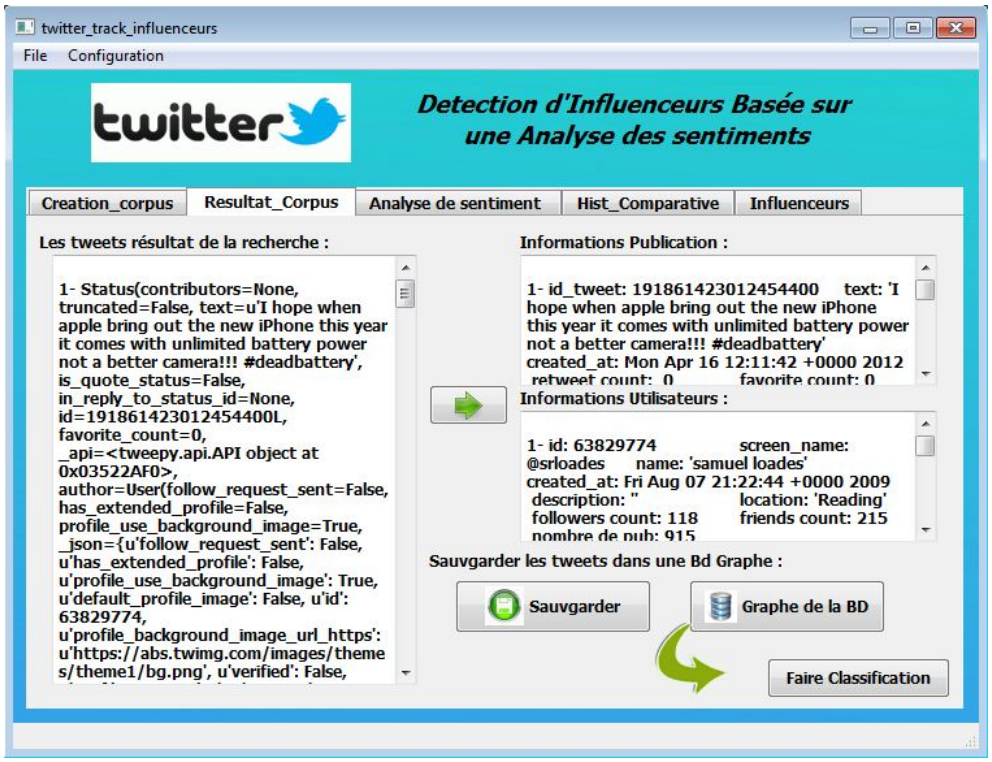

FIGURE III.4 – Interface d'affichage des résultat du corpus générer

A la fin de ce module, les données seront stockées sous forme de graphe dans la base de données. Et cela pour faciliter l'exploitation et la visualisation des différentes entités. La **figure III.6** représente notre base de données.

Légende :

Noeud Gris : Tweet.

Noeud rouge : User.

Flèches : Les relations entre Noeuds : « Tweeted », « Mentionned_IN », « Hashtags ».

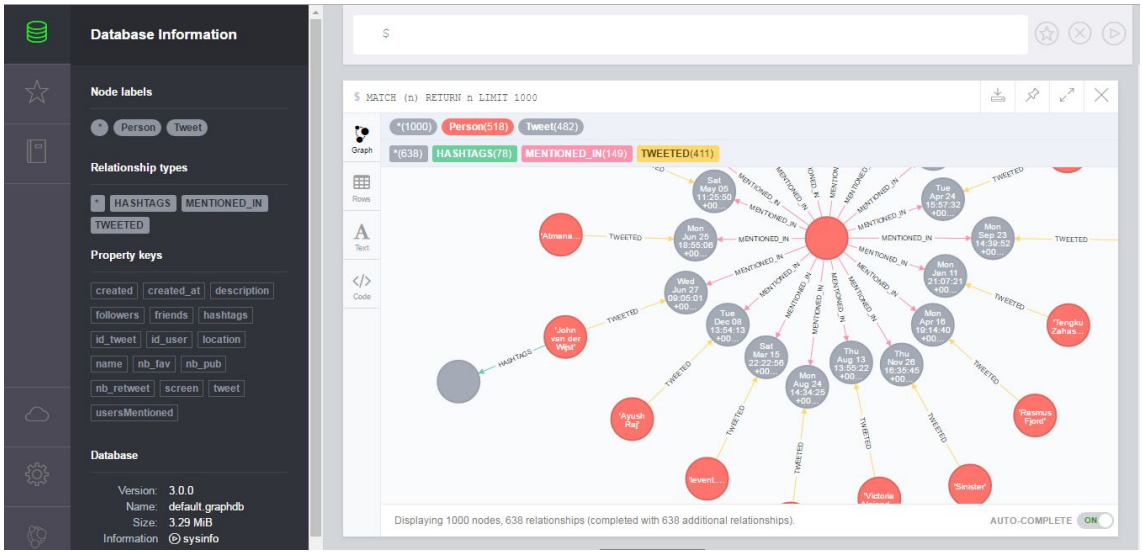

FIGURE III.5 – Représentation de la base de données graphe

III.4.2 Module d'analyse de sentiment

Après la phase de préparation soit terminé, les tweets sont donnés à la phase d'analyse de sentiments, afin de retourner en sortie des publications classées.Or, avant d'entamer cette étape l'utilisateur doit fixer le pourcentage d'apprentissage qui devrait être compris entre 60 et 80 % des données récoltés, ce paramètre permet de divisé le corpus d'étude en deux parties disjointe soit : les données d'apprentissage et les données de classification. Ce module passe par deux sous modules :

III.4.2.1 Sous module d'apprentissage

C'est principalement par apprentissage automatique que l'on tente de résoudre le problème de la catégorisation automatique de textes. Pour cela, cette phase est nécessaire pour annoter les données du corpus, en utilisent le package « sentiment » de R qui détermine la polarité, et donne aussi les résultats de l'émotion, elle classe l'émotion dans six catégories différentes, comme : la colère, le dégoût, la peur, la joie, la tristesse et la surprise.Le package sentiment de R comprend le classifieur bayésien pour identifier la classification.

III.4.2.2 Sous module de classification

A partir des données annotées, ce sous-module permet de classer le reste des données du corpus avec l'un des algorithmes de classification afin de générer trois classes de publications (positive, négative et neutre),nous avons choisi le classifieur naïf de Bayes, qui est l'une des méthodes les plus simples en apprentissage supervisé basée sur le théorème de Bayes permettant de calculer les probabilités conditionnelles. Dans un contexte général, ce théorème fournit une façon de calculer la probabilité conditionnelle d'une cause sachant la présence d'un effet, à partir de la probabilité conditionnelle de l'effet sachant la présence de la cause ainsi que des probabilités a priori de la

cause et de l'effet. On peut résumer son utilisation lorsqu'il est appliqué à la classification de textes ainsi : on cherche la classification qui maximise la probabilité d'observer les mots du document. Lors de la phase d'entraînement, le classificateur calcule les probabilités qu'un nouveau document appartienne à telle catégorie à partir de la proportion des documents d'entraînement appartenant à cette catégorie. Il calcule aussi la probabilité qu'un mot donné soit présent dans un texte, sachant que ce texte appartient à telle catégorie. Par la suite, quant un nouveau document doit être classé, on calcule les probabilités qu'il appartienne à chacune des catégories à l'aide de la règle de Bayes et des chiffres calculés à l'étape précédente [S, 2005] . Un des avantages de cette méthode est sa rapidité (même sur de très grandes bases de données). La **figure III.7** montre ce module d'analyse.

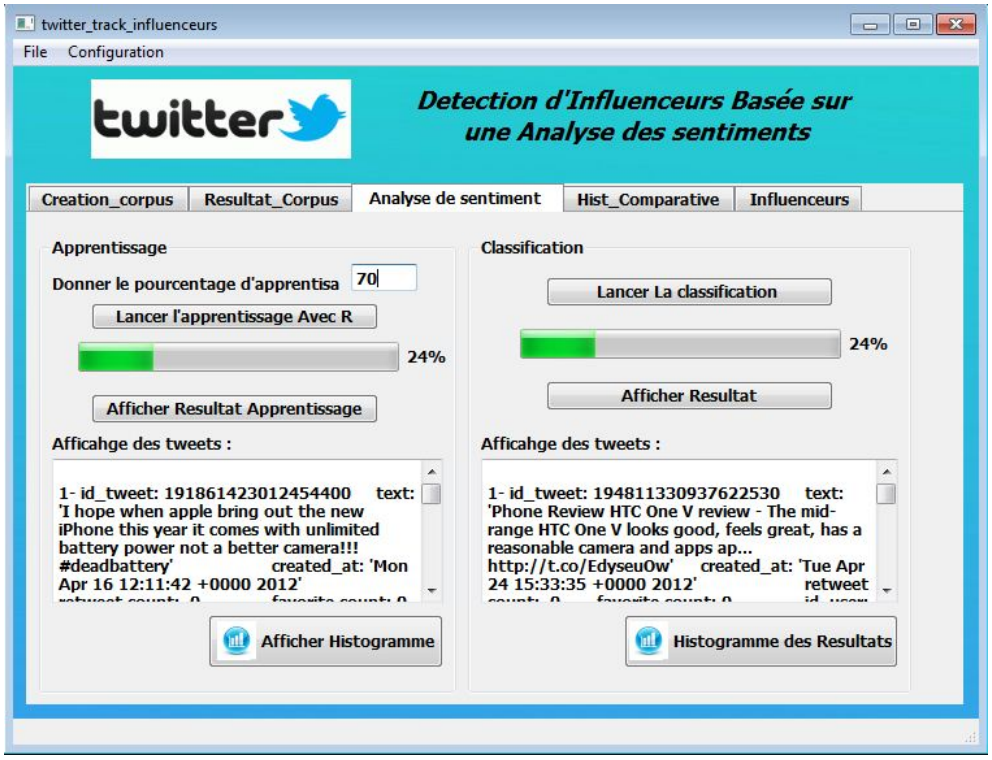

FIGURE III.6 – Interface du module d'analyse de sentiment

Le résultat comparatif des différentes marques après l'analyse de sentiments est présenté sous un histogramme comme le montre la **figure III.8** suivante :

III.4.3 Module de Détection d'influenceurs

Une fois les entités publications sont classées selon le sentiment que dégage le tweet, on s'intéresse à la source du critique pour identifier les utilisateurs qui sont considérés comme des potentiels influenceurs. Avant de débuter cette partie, l'utilisateur doit fixer le taux d'influence qui sera primordial pour l'identification des influenceurs. Dans un premier lieu, on doit attribuer pour chaque utilisateur le taux de popularité selon sa position dans la pyramide d'influence. Dans un second lieu, un ratio représentant l'impact de l'utilisateur dans le réseau social sera calculé. Ensuite, on

les additionne pour avoir à la fin le taux d'influence, qui permet de déterminer les potentiels influenceurs. La **figure III.9** présente ce module.

FIGURE III.7 – Interface du module de détection d'influenceurs

III.5 Présentation de la base de données graphe des influenceurs

Après la détection des potentiels influenceurs un graphe représentant les influenceurs et leur abonnées (followers) et abonnements (following) sera généré. Dans la base de données chaque nœud représente des utilisateurs influenceurs ou les abonnés et les abonnements. Voici un aperçu de notre base de données des influenceurs dans la **figure III.10**.

Légende :

Nœud Gris : Utilisateur influenceur.

Nœud Rouge : Follower.

Nœud Blue : Following.

flèches :Les relations entre Noeuds : « Follows », « Following ».

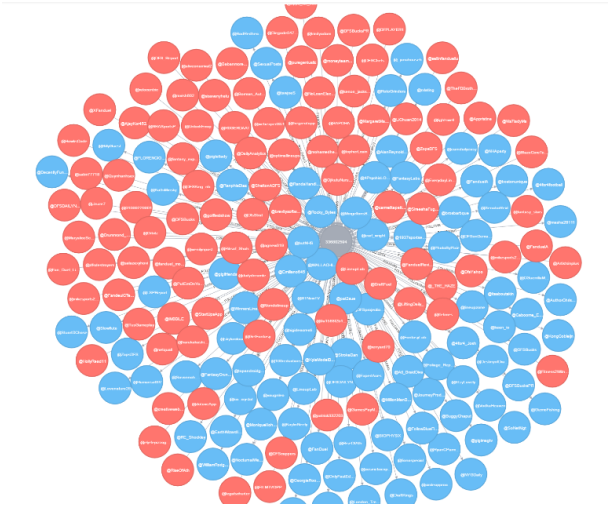

FIGURE III.8 – Représentation de la base de données des influenceurs

III.6 Conclusion

Nous avons décrit tout au long de ce chapitre l'implémentation de notre application, au départ nous avons présenté les outils utilisés durant notre réalisation, ainsi que la technique d'authentification utilisée par Twitter pour l'extraction des données du réseau, et aussi notre base de données graphe.

Conclusion Générale

Nous nous sommes attachés dans ce modeste travail à proposer une approche pour la détection des influenceurs dans les réseaux sociaux, en se basant sur l'analyse de sentiment du contenu des publications.

L'analyse de sentiment constitue une rapide évolution dans le domaine de la recherche, en particulier depuis l'émergence du Web 2.0 et les technologies connexes, comme les réseaux sociaux notamment Twitter, avec l'inconvénient qu'ils traitent les tweets sans prendre en considération la source du critique. De même, la détection d'influenceurs a été étudié dans diverses approches, mais à base des statistiques et non pas sur le contenu des publications. Notre travail a été guidé par une étude préalable des travaux existants traitant des problèmes de détection des influenceurs et d'analyse de sentiment.

Le présent document a été organisé en deux parties. Au début nous avons présenté les notions liées à la recherche d'information et au concept de l'influence dans le web social, ainsi que l'analyse de sentiment. Dans la suite du document, nous avons présenté notre étude conceptuelle de l'approche et de l'outil, suivie des différentes étapes de l'application que nous avons développé, qui permette de récupérer des données issues du réseau Twitter, à travers l'utilisation d'API, afin de les analyser pour identifier les potentiels influenceurs.

Au terme de ce travail, nous pouvons dire qu'il nous a permis de mieux consolider nos connaissances en informatique et de mettre en pratique les connaissances acquises durant notre cursus, et de plus acquérir un nouveau savoir dans le domaine de la recherche d'information.Il nous a permis également de :

- Nous initier dans le domaine de traitement de langage naturel, ainsi l'analyse de sentiment dans les réseaux sociaux.
- Manipuler de grand volume de données issue des réseaux sociaux comme Twitter.
- Consolider et développer nos compétences en programmation avec le langage Python et dans la modélisation avec UML.

Perspectives

A travers notre étude, nous pouvons envisagés d'autres voies possibles pouvant améliorer notre outil, nous proposons comme perspectives :

- Etablir un système de recommandation d'influenceurs.
- Développer une application web et mobile pour faciliter la portabilité et la flexibilité de l'outil pour les spécialistes en marketing.
- Exploitation d'autres réseaux sociaux.
- Intégrer une fonctionnalité entièrement automatique de construction de l'ontologie.
- Adapter notre approche pour intégrer d'autres techniques de classification, comme Support Vector Machine (SVM).

Références

Bibliographie

[ref, a] Définition influence. http://fr.wikipedia.org/wiki/Influence_(psychologie).

[ref, b] Définition marketing. http://www.toupie.org/Dictionnaire/Marketing.htm.

[ref, c] Influence. http://www.sensduclient.com/2007/05/bill-et-jeff-propos-de-clients.html.

[ref, d] les chiffres clés des réseaux sociaux. http://www.blogdumoderateur.com/50-chiffres-medias-sociaux-2016.

[ref, e] L'impact des influenceurs. http://www.nielsen.com/us/en/press-room/2012/nielsen-global-consumers-trust-in-earned-advertising-grows.html.

[ref, f] Neo4j. http://www.d-booker.fr/content/68-introduction-bases-de-donnees-neo4j.

[ref, g] Nombre utilisateurs dans les rs. http://www.alexitauzin.com/.

[ref, h] revolutions arabes au moyen-orient. http://www.huffingtonpost.com/2011/02/11/egypt-facebook-revolution-wael-ghonim_n_822078.html.

[ref, i] Statistique sur réseaux sociaux. http://www.internetlivestats.com.

[A et al., 2011] A, A., B, X., I, V., O, R., and R, P. (2011). Sentiment analysis of twitter datag.

[A, 2014] A, G. (2014). Diffusion de l'information dans les médias sociaux : Modélisation et analyse.

[A and P, 2010] A, P. and P, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining.

[A, 2007] A, T. (2007). Classification de texte et estimation probabiliste par machine à vecteurs de support.

[A.M and M, 2010] A.M, K. and M, H. (2010). Users of the world, unite ! the challenges and opportunities of social media. page 61.

[Augure,] Augure. *Guide pour la mise en place de votre stratégie d'influence*.

[B, 2010] B, L. (2010). Sentiment analysis and subjectivity. *Handbook of Natural LanguageProcessing, Second Edition*.

[B and L, 2004] B, P. and L, L. (2004). A sentimental education : sentiment analysis using subjectivity summarization based on minimum cuts.

[D, 2011] D, P. (2011). *Des textes communautaires à la recommandation*. PhD thesis, Université Pierre et Marie Curie.

[Donati, 1994] Donati (1994). *la perspective relationelle dans l'intervention réseaux : fondements théorique. », dans : L.Sanicola (dir.), l'intervention de réseaux*. Paris.

[E and K.B, 2002] E, B. and K.B, S. (2002). *Buzz Marketing : les stratégies du bouche à oreille, Editions d'Organisation*. Paris.

[E, 2012] E, K. (2012). *Approche de migration d'une base de données relationnelle vers une base de données NoSQL orientée colonne*. PhD thesis, UNIVERSITE DE YAOUNDE I.

[E and P.F, 1955] E, K. and P.F, L. (1955). The part played by people in the flow of mass communications. *New York : The Free Press*.

[E et al., 2012] E, V., L, B., J.P, G., and A, V. (2012). Le rôle et l'identification des leaders d'opinion dans les réseaux sociaux traditionnels et virtuels : controverses marketing et pistes de recherche.

[E and L, 2004] E, V. and L, F. (2004). Communiquer avec les leaders d'opinion en marketing, comment et avec quels médias. page 35.

[G.N and C,] G.N, F. and C, Tarquinio", t. . L. y. . . a. . P.

[I et al., 2011] I, O., C, M., J, L., and I, S. (2011). Overview of the trec-2011 microblog track. proceedings of the twentieth text retrieval conference.

[J, 2011] J, N. (2011). Marketing des arts et de la culture et e-commerce – le marketing viral.

[J.P and A, 2009] J.P, G. and A, V. (2009). *Utilisation de Twitter pour la recherche en marketing.*

[J.P and A, 2010] J.P, G. and A, V. (2010). Identification des leaders d'opinion sur internet : utilisation des données secondaires issues de twitter.

[K, 2013] K, P. (2013). Automatic ontology generation for sentiment analysis from twitter.

[K and D, 2015] K, V. and D, V. (2015). Building a custom sentiment analysis tool based on an ontology for twitter posts.

[L et al., 2011] L, B. J., L, T., and M, B. (2011). un modèle de recherche d'information sociale dans les microblogs : cas de twitter.

[L, 2012] L, H. (2012). *Not onlysql*. PhD thesis, Haute École de Gestion de Genève.

[M et al., 2011] M, T., J, B., M, T., K, V., and M, S. (2011). Lexicon-based methods for sentiment analysis.

[N, 2006] N, H. (2006). *Ontologie de domaine pour la modélisation du contexte en recherche d'information*. PhD thesis, Université Paul Sabatier.

[N.B,] N.B, E. Social network sites : Definition, history and scholarship. *Journal of Computer-Mediated Communication], year = 2008, pages =22*.

[S, 2005] S, R. (2005). Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés.

[T and M, 1997] T, M. and M, H. (1997). Machine learning.

[T, 2009] T, V. (2009). La course aux « followers ». *Le Monde*.