الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي والبحث العلمي

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

جامعة قاصدي مرباح – ورقلة

UNIVERSITE KASDI MERBAH - OUARGLA

FACULTE DES NOUVELLES TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

DEPARTEMENT D'INFORMATIQUE ET DES TECHNOLOGIES DE L'INFORMATION

THESE

POUR L'OBTENTION DU DIPLÔME DE DOCTORAT

Option : SYSTEMES ET RESEAUX INFORMATIQUE

Intitulé

Modélisation sémantique des intérêts des internautes dans les réseaux sociaux

Présentée Par : BENKHELIFA Randa

Membres du Jury :

Dr. BENSAYAH Abdallah	Université de Ouargla	Président
Pr. LAALLAM Fatima Zohra	Université de Ouargla	Directrice de thèse
Pr. KAZAR Okba	Université de Biskra	Examinateur
Pr. KHERFI Mohammed Lamine	Université de Ouargla	Examinateur

Année universitaire : 2018/2019

Remerciements

*Je remercie **ALLAH** de m'avoir donné la force, le courage et la patience pour terminer ce travail.*

*J'aimerai aussi adresser mes plus sincères gratitudes à ma directrice de thèse, **Pr. LAALLAM Fatima Zohra**,*
Merci pour votre soutien, vos conseils, pour votre esprit de synthèse et votre érudition qui représentent un exemple à suivre. Merci pour votre confiance ininterrompue en mes compétences scientifiques et humaines.

*J'exprime également toute ma gratitude au **Pr. BISKRI Ismail** de l'université du Québec à Trois-Riviéres, Canada. Un homme de science m'a fait l'honneur de travailler en sa compagnie. IL a pris le temps de m'écouter et de discuter avec moi. Ses remarques m'ont permis d'envisager mon travail sous un autre angle. Merci du fond du cœur pour tout cela.*

*Je remercie aussi très vivement **Mme DRID Touria**, docteur à l'université d'Ouargla pour la correction de mes articles en langue anglaise.*

*Mes remerciements vont aussi à **Mme ESMA Aimeur**, Professeure à l'Université de Montréal pour ses conseils et son accueil chaleureux au sein de son laboratoire pendant mon stage au Canada.*

Je tiens aussi particulièrement à exprimer mes vifs et sincères remerciements au ***Pr. Mohammed Lamine KHERFI*** *de l'université de Ouargla pour l'intérêt et l'attention qu'il a accordé à ce travail de thèse et d'avoir accepté de faire partie de ce jury en tant qu'examinateur.*

Je remercie très vivement ***Mr. BENSAYAH Abdallah****, docteur à l'université de Ouargla d'avoir accepté d'être le Président du Jury.*

Ces mêmes remerciements s'adressent au ***Pr. KAZAR Okba*** *professeur à l'université de Biskra d'avoir accepté de faire partie de ce jury en tant qu'examinateur.*

Merci à tous les professeurs qui m'ont donné l'envie d'apprendre, la curiosité intellectuelle, et le goût de la recherche.

Merci à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce travail.

Et enfin, je tiens à remercier chaleureusement l'ensemble de mes collègues et ami(e)s du laboratoire LINATI.

Résumé

Les sites des réseaux sociaux tels que Facebook et Twitter sont en croissance exponentielle ; ils génèrent des quantités considérables de données personnelles et sociales sur les utilisateurs et leurs interactions. Ces données représentent une source d'informations importante pour plusieurs applications telles que les systèmes de recommandation, la détection des communautés et le marketing. De telles applications nécessitent des informations utiles sur les données personnelles et le contenu social de l'utilisateur. Cependant, la structure actuelle des réseaux sociaux ne permet pas d'extraire de manière simple et rapide ces informations. Et ce, malgré les efforts considérables déployés par la communauté du Web sémantique afin de résoudre ce problème, en essayant de représenter l'utilisateur de la manière la plus précise possible. Néanmoins, ces modèles sémantiques sont incapables de donner un sens au contenu. De ce fait, il faut exploiter le contenu textuel afin d'enrichir le profil utilisateur.

Dans cette thèse, nous présentons un modèle générique qui combine deux disciplines différentes (le text mining et les modèles sémantiques). Le modèle proposé intègre l'approche du text mining dans un modèle sémantique afin d'enrichir le profil utilisateur en classifiant son contenu textuel. Le but est de: (1) regrouper le contenu social par catégories selon une approche descendante, (2) déduire et modéliser les intérêts dynamiques temporelles de l'utilisateur, (3) regrouper les utilisateurs qui ont des intérêts similaires par catégories, (4) mettre au point un mécanisme pour interroger les données du système.

Après cela, nous évaluons le modèle proposé en construisant une application sociale qui utilise un grand ensemble de données Facebook. Ce modèle peut être utilisé comme structure de base pour plusieurs applications telles que les systèmes de recommandation.

Mots clés : Modélisation sémantique, Text mining, Classification du texte, Réseaux sociaux, Intérêts d'utilisateur, Web sémantique, Profilage.

Abstract

Online social networks such as Facebook and Twitter are in exponential growing, they generate considerable amounts of personal and social data about users and their interactions. This data is an important source of information for several applications such as recommendation systems, community detection and marketing. Such applications require useful information about the personal data and the social content of the user. However, the current structure of social networks does not allow to extract this information in a simply way. Despite, major efforts have emerged from the semantic web community to address this problem trying to represent the user as accurately as possible. These Semantic Models are unable to give a sense to the user-generated content. For this, more mining and sense-making need to be done on the content, to enrich the user profile.

In this thesis, we introduce a generic model which combines two disciplines (Text Mining and Semantic models). The proposed model incorporates the text mining approach into a semantic model to enrich the user profile by including information on user's posts classifications. In an attempt to: (1) group online contents into topics under a top-down approach, (2) infer and model the temporal dynamic interests of users, (3) group users with similar interests and preferences, (4) attribute a mechanism for querying the system data.

We evaluate the proposed model by building a social application that uses a large Facebook dataset. This model can be used as a basic structure for several applications such as recommendation systems.

Keywords: Semantic modeling, Text mining, Text classification, Online social networks, User interests, Semantic Web, Profiling.

ملخــص

تشهد مواقع الشبكات الاجتماعية مثل فيسبوك و تويتر تزايد مستمر، فهي تولد كميات كبيرة من البيانات الشخصية والاجتماعية حول المستخدمين وتفاعلاتهم. تمثل هذه البيانات مصدرًا مهمًا للمعلومات بالنسبة للعديد من التطبيقات مثل أنظمة الإقتراح و اكتشاف التجمعات و التسويق. تتطلب مثل هذه التطبيقات معلومات مفيدة حول البيانات الشخصية والمحتوى الاجتماعي للمستخدم. إلا أن البنية الحالية للشبكات الاجتماعية لا تسمح لاستخراج هذه المعلومات ببساطة وبسرعة. على الرغم من الجهود الكبيرة التي يبذلها مجتمع الويب الدلالي لحل هذه المشكلة ، بمحاولته لتمثيل المستخدم بأكثر دقة ممكنة. نتيجة لذلك يجب استغلال المحتوى النصي لإثراء بروفيل المستخدم. في هذه الأطروحة، نقدم نموذجًا عامًا يجمع بين تخصصين (تحديدًا تعدين النص والنماذج الدلالية). النموذج المقترح يدمج منهج تعدين النص في نموذج دلالي لإثراء بروفيل المستخدم من خلال تصنيف محتواه النصي. و ذلك من أجل: (1) تصنيف المحتوى الاجتماعي حسب الفئات باتباع نهج الهرم ، (2) استنتاج ونمذجة اهتمامات المستخدمين الديناميكية بالنسبة للزمن، (3) تصنيف المستخدمين الذين يملكون اهتمامات مماثلة إلى فئات (4) تطوير آلية للاستعلام عن بيانات النظام. بعد ذلك، نقوم بتقييم النموذج المقترح من خلال بناء تطبيق اجتماعي يستخدم مجموعة كبيرة من بيانات فيسبوك. يمكن استخدام هذا النموذج كهيكل أساسي للعديد من التطبيقات مثل أنظمة الإقتراح.

الكلمات المفتاحية: النمذجة الدلالية، تعدين النص، تصنيف النصوص، الشبكات الاجتماعية، اهتمامات المستخدم، الويب الدلالي، بروفيل.

SOMMAIRE

Sommaire

Remerciements II
Résumé IV
Sommaire VIII
Liste des figures XI
Liste des tableaux XII
Introduction générale 1
Contexte de travail et problématique 1
Contribution 2
Organisation de la thèse 3
Chapitre 1 : Etat de l'art 6
1.1. Introduction 6
1.2. Les plateformes sociales et les réseaux sociaux 6
1.2.1. Les plates-formes du Web social 6
1.2.1.1. Communautés virtuelles 7
1.2.1.2. Classification des plates-formes sociales selon le contenu partagé 13
1.2.1.3. Participation de l'utilisateur dans les plateformes sociales 15
1.2.2. Les réseaux sociaux 15
1.2.2.1. Historique des réseaux sociaux 15
1.2.2.2. Définitions des réseaux sociaux 17
1.2.2.3. Les catégories principales des réseaux sociaux selon leurs objectifs et usages 17
1.2.2.4. Les réseaux sociaux comme source de données 18
1.2.3. Les caractéristiques du contenu social 18
1.2.4. Les médias sociaux 19
1.3. Le Web Social Sémantique 19
1.3.1. Introduction 19
1.3.2. Le web Sémantique 20
1.3.3. Le Web social sémantique 21
1.3.4. Les vocabulaires du Web social sémantique 22
1.3.4.1. Ontologies pour modéliser les profiles utilisateurs 22
1.3.4.2. Les modèles sémantiques pour capturer le contenu partagé dans les RS 23
1.4. Prétraitement et catégorisation du texte social 26
1.4.1. Le Data Mining 26

1.4.2. L'apprentissage automatique 27
1.4.2.1. Type d'apprentissage 27
1.4.2.2. Le deep learning 28
1.4.3. Le Text Mining 29
1.4.4. La catégorisation de texte 29
1.4.5. Traitement du langage naturel 30
1.4.5.1. La représentation de texte 31
1.4.5.2. Term weighting scheme (TWS) 32
1.4.6. La classification des textes issus des réseaux sociaux 33
1.5. La modélisation des intérêts des utilisateurs et catégorisation des utilisateurs 35
1.5.1. Déduction et modélisation des intérêts des utilisateurs 35
1.5.2. Catégorisation des utilisateurs 37
1.5.3. Conclusion 37
Chapitre 2 : Méthodologie 39
2.1. Introduction 39
2.2. L'architecture du modèle UCC 39
2.3. L'ontologie UCC 41
2.3.1. Modélisation des profils utilisateurs 41
2.3.2. Enrichissement des profils utilisateurs 43
2.3.3. L'implémentation de l'ontologie 46
2.4. Arbre hiérarchique de catégorisation du contenu 46
2.5. La déduction des intérêts des utilisateurs 47
2.6. La catégorisation des utilisateurs 49
2.6.1. Le regroupement d'utilisateurs en fonction de la similarité d'intérêts avec les utilisateurs-cœur 49
2.6.2. Le regroupement des utilisateurs en fonction d'une catégorie d'intérêt 51
2.7. Exploitation des informations démographiques dans la classification du texte social 53
2.8. Conclusion 58
Chapitre 3 : Expérimentation et discussion 60
3.1. Introduction 60
3.2. Collecte et construction de données d'expérimentation 60
3.3. Métriques d'évaluation des modèles d'apprentissage automatique 62

3.4. Analyse quantitative du lien entre les attributs démographiques de l'utilisateur et le texte généré par l'utilisateur 65
3.5. Prétraitement de texte et les « term weighting schemes » utilisés 67
3.6. Résultats et discussion 67
3.6.1. Résultat des classifications 67
3.6.2. Analyser la dynamique des intérêts des utilisateurs 77
3.6.3. La similarité d'intérêt entre les utilisateurs 79
3.6.4. Recherche des données par l'ontologie UCC 80
3.7. Comparaison entre le modèle UCC et les autres modèles 82
3.8. Conclusion 84
Conclusions générales et perspectives 86
Champs d'application 87
Limitations 88
Perspectives 88
Références 90

Liste des figures

Figure 1.1 : Taxonomie des plates-formes sociales : principaux piliers et dimensions correspondantes Stan, (2011). 9
Figure 2.1: Le processus d'utilisation du modèle UCC. 39
Figure 2.2: L'ontologie fondamentale du SIOC Breslin et al., (2005). 41
Figure 2.3: L'ontologie UCC 45
Figure 2.4 : Une prise d'écran de l'ontologie UCC. 46
Figure 2.5 : un exemple d'un arbre hiérarchique de catégorisation du contenu. 47
Figure 2.6: un exemple d'un arbre d'intérêt d'un utilisateur User-A en novembre 2015. 48
Figure 2.7: un exemple de regroupement des utilisateurs en fonction de la similarité d'intérêts avec les utilisateurs-cœur 51
Figure 2.8 : un exemple de regroupement des utilisateurs en fonction d'une catégorie d'intérêt 53
Figure 2.9: une illustration pour expliquer la relation entre attributs et intérêts des utilisateurs Sujet des intérêts et intérêts similaires. 54
Figure 2.10:le processus de la classification de l'approche basée sur le contenu et les attributs démographique de l'utilisateur. 55
Figure 3.1: les distributions d'attributs démographiques de 8 catégories de la classification générale (dataset-1). 65
Figure 3.2: les résultats de F-mesure obtenus en utilisant le classifier ANN sur les versions de données (V1 à V7). 67
Figure 3.3: les résultats de F-mesure obtenus en utilisant le classifier KNN sur les versions de données (V1 à V7). 68
Figure 3.4: les résultats de F-mesure obtenus en utilisant le classifier NB sur les versions de données (V1 à V7). 68
Figure 3.5: les résultats de F-mesure obtenus en utilisant le classifier DT sur les versions de données (V1 à V7). 69
Figure 3.6: les résultats de F-mesure obtenus en utilisant le classifier DR sur les versions de données (V1 à V7). 69
Figure 3.7: les résultats de F-mesure obtenus en utilisant le classifier SVM sur les versions de données (V1 à V7). 70
Figure 3.8: la CCHT construite de l'apprentissage pour l'application. 77
Figure 3.9 : la similarité des intérêts de l'utilisateur User_A entre le mois en cours et les mois précédents. 78
Figure 3.10: évolution périodique des 5 principales catégories d'intérêts de l'utilisateur User-A 79
Figure 3.11: la similarité entre l'utilisateur User-A et 3 autres utilisateurs 80

Liste des tableaux

Tableau 1.1 : Un panorama des plateformes sociales en ligne les plus populaires pour chaque catégorie Stan, (2011). 14
Tableau 1.2 : Des différents schémas de pondération..... 33
Tableau 2.1 : Le fichier des utilisateurs. 52
Tableau 2.2 : La construction des groupes d'utilisateurs 52
Tableau 3.1 : Les statistiques de données de dataset-1pour la classification Général. 61
Tableau 3.2 :La distribution des catégories pour la classification News (dataset-1). 61
Tableau 3.3 : Liste des attributs démographiques. 61
Tableau 3.4: La distribution des catégories pour la classification Générale (dataset-2). 62
Tableau 3.5 :La distribution des catégories pour la classification News (dataset-2). 62
Tableau 3.6:les attributs démographique de l'utilisateur sélectionné dans chaque version de jeu de données (V1 à V7), Precision (P), Recall (R), F-Measure (F), and Accuracy (A), utilisant la validation croisée de 10 fold. 72
Tableau 3.7: Les résultats obtenus par SVM avec les versions de jeu de données V7 et V1 (baseline) pour chaque catégorie. 76
Tableau 3.8 : Les résultat d'apprentissage utilisant SVM pour la classification News 77
Tableau 3.9: comparaison entre le modèle UCC et les autres modelés de l'état de l'art..... 83

Introduction générale

Introduction générale

Contexte de travail et problématique

Il y a quelques années, le passe-temps préféré des gens était de regarder la télévision. Aujourd'hui, les ordinateurs, les tablettes et les Smartphones ont pris la relève. Les utilisateurs et plus spécialement les jeunes et les ados « la génération Web » passent des heures et des heures devant ces écrans, en navigant sur le Web et particulièrement sur les sites interactifs, qui ont vu le jour avec la naissance du web 2.0. Ces sites permettent à leurs utilisateurs de communiquer, de partager l'information, d'établir des relations personnelles et de s'exprimer librement. Cela se passe à travers de multiple-plateforme comme les wikis, les forums, les blogs et surtout les réseaux sociaux qui ont bouleversé les usages d'Internet depuis leurs apparitions.

Aujourd'hui, les réseaux sociaux ne sont plus des simples sites web, mais un outil populaire de communication entre les utilisateurs. À chaque minute qui passe, ils génèrent des volumes considérables de données sur l'utilisateur et ses interactions. Les utilisateurs peuvent facilement créer, accéder et partager des contenus, des news, des opinions et des informations en général. Habituellement, ces utilisateurs ne se soucient pas de l'orthographe et de la construction grammaticale exacte d'une phrase. Ils utilisent le langage SMS (langage texto), des textes informels « des nouvelles formes de communication écrite » Véronis, and Guimier. (2006). Les réseaux sociaux sont un phénomène social qui a un très grand nombre d'utilisateurs à travers le monde.

Les données des réseaux sociaux sont de grande variété, et elles représentent des sources importantes d'informations. Mais malheureusement, l'accès à ces données devient de plus en plus difficile. Cependant, organiser, modéliser et accéder rapidement à l'information devient alors un réel challenge face à cette grande masse de données.

Plusieurs études portent sur la modélisation des caractéristiques des utilisateurs et des interactions dans les réseaux sociaux. Des efforts importants ont été déployés par la communauté du Web sémantique pour résoudre ce problème en essayant de représenter l'utilisateur de la manière la plus *précise* possible. En représentant des vocabulaires dans un format standard, tels que RDF et OWL. Ces vocabulaires ont été développés pour modéliser sémantiquement les utilisateurs et leur contexte social, y compris les vocabulaires FOAF *(Friend of a Friend),* and

SIOC *(Semantically Interlinked Online Communities).*

Les données sociales sont décrites sous forme RDF (*Resource Description Framework*), un graphe typé qui fournit une représentation plus puissante et plus riche des réseaux sociaux par rapport à l'analyse classique des graphes des réseaux sociaux. Malgré la puissance de ces ontologies en modélisant les caractéristiques et les interactions des utilisateurs au sein des communautés en ligne, elles sont incapables de donner un sens au contenu. Elles ne peuvent pas modéliser des aspects dynamiques tels que les intérêts des utilisateurs, ni de regrouper les utilisateurs isolés discutant les mêmes sujets ou bien capturer les mêmes messages traitant le même sujet. Cependant, il faut ajouter du sens au contenu. Le text mining et le traitement du langage naturel sont les meilleures disciplines pour traiter ce problème. Plusieurs recherches ont tenté d'extraire un sens du texte libre et d'apporter des améliorations utiles dans plusieurs applications.

Dans cette thèse, nous nous sommes particulièrement intéressés aux utilisateurs des réseaux sociaux, ainsi que les textes postés par eux. Ces textes non structurés peuvent être considérés comme des sources importantes d'informations. Il est donc nécessaire de disposer d'outils informatique afin de mettre en place des systèmes capables de structurer les informations issues des réseaux sociaux et de traiter directement les textes exprimés en langage naturel pour pouvoir organiser et accéder facilement à l'information et pour permettre ensuite l'exploitation de ces connaissances afin de comprendre le comportement des internautes et d'offrir des statistiques en se basant sur des outils et techniques adaptés. Par conséquent, et pour mettre en œuvre ces systèmes, il faut savoir :

- Comment peut-on représenter (réordonner) les informations brutes issues des réseaux sociaux d'une façon à être utilisable, facile à trouver et exploitable par la suite ?
- Comment peut-on catégoriser le contenu textuel issu des réseaux sociaux tout en profitant de la nature de ces textes ?
- Comment peut-on déduire et modéliser les intérêts dynamique et temporel des utilisateurs du contenu textuel partagé par eux ?
- Et est-ce-que la classification du contenu peut enduire à une classification des utilisateurs ?

Contribution

Notre objectif est de trouver une structure « un modèle sémantique » qui permet d'extraire automatiquement des informations et des connaissances cachées sur les utilisateurs et le contenu textuel écrit en langage naturel par eux dans les réseaux sociaux.

Afin de pouvoir répondre aux questions précédentes et de réaliser notre modèle :

- La proposition d'un modèle générique qui réunit deux techniques différentes (le text mining et les modèles sémantiques) afin de décrire les utilisateurs et d'analyser leur contenu partagé à partir de réseaux sociaux. Nous choisissons une modélisation basée sur les ontologies.
- L'attribution des classifications automatiques au contenu textuel en plusieurs niveaux en se basant sur les algorithmes d'apprentissage automatique sous des catégories choisit par l'utilisateur final de notre modèle.
- La proposition d'une nouvelle approche basée sur le contenu et les attributs démographiques afin d'améliorer les performances de la classification des textes sociaux.
- L'inférence et la modélisation des intérêts dynamiques et temporels de l'utilisateur avec la création d'un arbre d'intérêts d'utilisateur scorée en utilisant une approche hiérarchique.
- La proposition de deux approches permettant de classifier les utilisateurs en fonction de leurs intérêts (regroupement des utilisateurs selon une approche de similarité d'intérêts avec les utilisateurs-cœur et regroupement des utilisateurs selon une approche de catégorie d'intérêts).
- La proposition d'un mécanisme configurable pour interroger le modèle proposé.

Notre objectif se focalise sur l'utilisation de modèles ontologiques pour représenter et analyser sémantiquement les réseaux sociaux. Un des objectifs scientifiques est d'améliorer l'analyse sémantique des réseaux sociaux en réalisant des inférences sur ces modèles proposés de ces réseaux grâce à l'utilisation d'ontologies dédiées. Cette nouvelle fonctionnalité va permettre dans un premier temps de détecter plus facilement des communautés d'intérêts et dans un second temps, grâce à la conception d'algorithmes adaptés permettant de suivre l'activité de ces communautés, de concevoir des services à valeur ajoutée grâce aux connaissances acquises dans l'étape d'analyse.

Les résultats obtenus à la fin de cette thèse permettront par la suite d'offrir une vue globale sur les utilisateurs, leurs centres d'intérêts, et leurs regroupements. Ceci ouvre une grande porte devant une variété de spécialistes comme les sociologues, les politiciens, etc.

Organisation de la thèse

Cette thèse est organisée en trois chapitres.

- **Chapitre 1 :** Le premier chapitre présente le contexte général de notre travail : La première partie présente l'aspect général des plateformes du Web social, et leurs caractéristiques. Ensuite, nous introduisons les réseaux sociaux. La deuxième partie présente l'état de l'art relatif à l'utilisation des technologies du web sémantique pour modéliser les profils utilisateurs. La troisième Partie porte sur l'état de l'art présenté concernant les techniques du

data mining, text mining, et la classification du texte pour organiser et catégoriser les données textuelles issues des réseaux sociaux. La quatrième et dernière partie du chapitre donne un aperçu des travaux réalisés pour déduire et modéliser les intérêts des utilisateurs et pour catégoriser les utilisateurs selon leurs intérêts.

- **Chapitre 2 :** Ce chapitre détaille notre contribution. Il présente notre modèle générique proposé qui permet d'enrichir le profil utilisateur (modèles sémantiques) par la classification de contenu textuel. Il montre aussi la proposition d'une nouvelle approche pour améliorer la performance de la classification de ce contenu textuel.
- **Chapitre 3 :** le troisième chapitre détaille l'expérimentation, les résultats obtenus, et une discussion de ces résultats. L'évaluation est réalisée sur un grand ensemble de données Facebook.

En fin de cette thèse, nous discutons des implications que peuvent avoir les propositions présentées dans l'amélioration des mécanismes associés dans les systèmes d'information, ainsi que de nombreuses pistes de recherche futures.

Chapitre I

Etat de l'art

1. Chapitre 1 : Etat de l'art

1.1. Introduction

Le développement considérable subi par les technologies au cours des dernières décennies a conduit à l'émergence des applications Internet sur la base logiciels et services conçus pour améliorer la collaboration en ligne pour le grand public, et ce, sous forme de blogs, wikis, sites de partage vidéo, et les réseaux sociaux. Les réseaux sociaux sont devenus l'une des principales plates-formes actuelles d'amitié et de partage des intérêts.

Ce chapitre est divisé en quatre parties. La première partie représente les plateformes sociales, et leurs caractéristiques, et plus spécialement les réseaux sociaux. La deuxième partie explore les ontologies développées pour modéliser sémantiquement les profils des utilisateurs dans le contexte des médias sociaux et des communautés en ligne. La troisième explore les techniques de text mining, afin d'analyser et de classifier les textes sociaux, en donnant un aperçu sur les travaux relatifs à ces disciplines. Enfin, la quatrième partie discute les études existantes liées aux propriétés des intérêts des utilisateurs et le regroupement des utilisateurs dans des réseaux sociaux.

1.2. Les plateformes sociales et les réseaux sociaux

Les blogs, les wikis, les sites de partage vidéo, et les réseaux sociaux, se regroupent tous sous le terme plateformes sociales. Ces plateformes sociales en ligne font partie d'une tendance technologique du Web 2.0. Ce dernier a mis le Web en valeur avec plus d'interactions et de collaboration. Depuis sa naissance, les plates-formes sociales ont fait des utilisateurs plus actifs en les permettant d'avoir des listes d'amis, des profils utilisateur et des partages de contenu. En conséquent, elles peuvent être considérées comme des communautés virtuelles, dans lesquelles les utilisateurs partagent leurs propres informations démographiques et sociales, communiquent, partagent de contenu, réagissent et commentent, ou tout simplement s'interagir entre eux.

Dans cette partie, nous présentons les caractéristiques générales des plates-formes sociales, en définissant un modèle général pour ces plates-formes en identifiant les blocs fondamentaux de leurs constructions, et plus spécialement, en détaillant les réseaux sociaux.

1.2.1. Les plates-formes du Web social

Une plate-forme sociale est généralement forgée par une communauté virtuelle des membres. Le concept de Communautés Virtuelles (CV) a ses origines dans la définition des systèmes composés d'agents autonomes, capables d'interagir et de collaborer pour atteindre un objectif

Camarinha-Matos and Afsarmanesh, (2004), Maret et al., (2004). Cependant, avec l'émergence des premières plates-formes sociales, ce concept a également été utilisé pour définir la communauté en ligne sous-jacente des plates-formes sociales Wellman, (1996).

1.2.1.1. Communautés virtuelles

Dans cette section, nous discutons d'abord le concept de communautés virtuelles ainsi que ses principaux piliers. Ensuite, nous examinons la nature des contenus produits par ces communautés. Et, en dernier lieus, nous discutons les défis de la gestion et le traitement de l'information sociale.

a. Les communautés en ligne

Le mot communauté est un terme qui évolue à travers le temps avec le changement des technologies et des comportements humains. Le concept de communauté (en sociologie) est défini comme un groupe d'interaction entre un ensemble de gens, vivant dans un endroit commun et organisé autour de valeurs communes. Depuis l'apparence d'Internet, le concept de communauté n'a plus de limites géographiques. Où, les gens peuvent maintenant se rassembler virtuellement dans les communautés en ligne, formant des communautés virtuelles où ils peuvent partager des intérêts communs indépendamment de leur emplacement physique.

- Historiquement
  - ✓ Au début des années 70, les technologies d'Internet dans ses premières formes telles que les systèmes de babillard (BBS), Internet Relay Chat (IRC), Usenet et les forums ont été fortement liées à la notion de communauté.
  - ✓ En 1985, le Well (The Whole Earth 'Lectronic Link) Rheingold, (2000) est apparaît comme la première véritable communauté en ligne. C'est un système informatisé de conférence qui permet aux gens dans le monde entier à mener des conversations publiques et échanger du courrier électronique privé (e-mail).
  - ✓ À ce stade précoce, les communautés en ligne ont été utilisées pour décrire les personnes qui utilisent Internet en faisant des discussions en ligne sur un espace de communication appelé le cyberespace.
  - ✓ Le développement rapide des communautés virtuelles en ligne en vraiment eu lieu après le dot.com en 2002, où l'Internet et le Web a été réinventé avec le label Web 2.0. Depuis cette période, les interactions sociales et le contenu généré par l'utilisateur dans des espaces virtuels ont caractérisé la principale forme d'interactions.

- ✓ Les réseaux sociaux, tels que Facebook ou LinkedIn et plus récemment Twitter sont généralement connus comme la partie la plus réussie et visible des communautés en ligne.

b. Les composantes des communautés virtuelles

Les communautés virtuelles se composent des utilisateurs et des objets sociaux représentant les intermédiaires des interactions entre les utilisateurs.

Dans les sous-sections suivantes, nous décrivons les caractéristiques des utilisateurs, comme le profil utilisateur, les activités, et les connexions, entre les utilisateurs au sein des communautés virtuelles dans les plates-formes sociales.

- ❖ **Principaux piliers des plates-formes sociales**

  Dans les plates-formes sociales, la communauté virtuelle est formée d'utilisateurs et de ceux que l'on appelle les objets sociaux représentant les intermédiations. Les objets sociaux sont communément appelés les contenus générés par utilisateurs (User-Generated Content) dans le cas du Web 2.0. Ils peuvent être :

  - Ressources (représentant des objets partagés, comme des photos, des vidéos, des pages Web, mais peut aussi être un objet physique ayant une identité numérique).
  - Annotations (teneur sociale qui décrivent une ressource, tels que les balises sociales, microbloging, etc.)

Le Web 2.0 a introduit une nouvelle liberté dans la relation des utilisateurs avec le Web grâce à de nouveaux processus de communication en ligne. L'utilisateur est maintenant en mesure de créer facilement du contenu et de le rendre disponible pour le monde entier, et aussi annoter, ajouter des commentaires ou évaluer le contenu existant. Certains utilisateurs passent beaucoup de temps sur les plateformes sociales pour échanger des informations au sein de leurs communautés sociales. Par conséquent, les gens sont organisés dans des communautés autour de topiques et intérêts spécifiques. Les interactions au sein d'une plateforme sociale peuvent être très disparates et souvent parlant de sujets inattendu, mais ils restent très intéressants sur le point de vue utilisateurs.

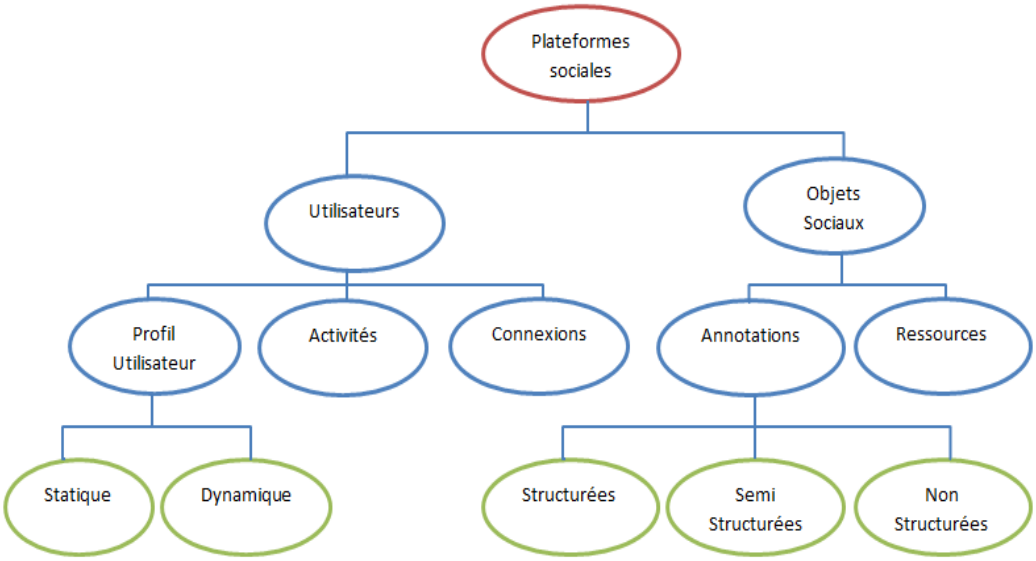

Figure 1.1 : Taxonomie des plates-formes sociales : principaux piliers et dimensions correspondantes Stan, (2011).

- **Utilisateurs**

  Toute plate-forme sociale est composée d'un ensemble d'utilisateurs, et dans laquelle, tout utilisateur est caractérisé par un profil, des activités et des connexions.

  ✓ **Le profil utilisateur :** un profil comprend généralement des informations statiques personnelles, telles que le nom, l'email et l'adresse, ainsi que des informations plus dynamiques comme les intérêts, les préférences et les besoins de l'utilisateur. Un profil utilisateur joue un rôle essentiel dans les communautés en ligne. Généralement, les profils utilisateur sont différents d'une application à une autre, car les utilisateurs se présentent différemment, en fonction de la population cible d'une application donnée (qui sont quelques fois de très spécifique). A la plupart du temps, dans un profil utilisateur au sein d'un réseau social, tout utilisateur a des connexions d'amitié, ces connexions sont formées en fonction des loisirs et des activités sociales. Dans une plateforme ciblée pour le réseautage professionnel, un profil contient des informations sur les intérêts professionnels de l'utilisateur, tels que les compétences techniques, l'expérience et les plans futurs. Plus généralement, un utilisateur peut être membre de plusieurs plates-formes sociales et ont différentes facettes d'identité dans chacune d'eux. Une facette d'identité est un profil ciblé à une communauté bien définie.

  ✓ **Activités** : Une autre dimension des utilisateurs est représentée par les activités qu'ils exercent dans une plate-forme sociale. Cela inclut le partage de contenu, le téléchargement des documents multimédias et la description du contenu (tels que

le marquage de photos). L'augmentation semble accélérer à mesure que de plus en plus d'applications sont disponibles.

- ✓ **Connexions sociales** : la troisième dimension des utilisateurs au sein d'une plateforme sociale est représentée par les liens sociaux qu'ils établissent avec les autres utilisateurs. Les utilisateurs des plateformes sociales sont généralement connectés à des communautés différentes, appartenant à différentes sphères sociales (par exemple amis, famille, collègues de travail, etc.). Deux grandes catégories de connexions peuvent être distinguées :
  a. Connexions indirectes : cette catégorie de connexions se réfère à la situation où un accord mutuel est nécessaire afin d'établir le lien social. Un bon exemple est le cas des amitiés dans un réseau social. L'objectif de ces connexions est de fournir aux gens une sorte de conscience sociale sur les activités, et les plans de leurs amis. Ceci peut être réalisé en visitant leur profil ou en recevant des notifications s'il y a une mise à jour dans leur profile.
  b. Connexions directes : une connexion directe est une forme plus récente d'établir des liens sociaux. Cela signifie le fait qu'un utilisateur peut vouloir recevoir une notification d'un autre, mais pas vice versa. Ce type de connexion est destiné à suivre son activité et de ne pas exprimer une amitié dans la vie réelle.

- **Objets sociaux**

  Comme mentionné précédemment, le contenu partagé influence les interactions entre les utilisateurs. Cette observation a donné naissance au principe de la notion objet centré. Un objet est un point focal d'intérêt commun, physique et / ou numérique. Certains objets sont la source des interactions et des attracteurs de l'attention collective de conversation. Ils constituent un support de conversation. Dans les objets de nature numériques réels sont ceux principalement des multimédias comme des articles (Wordpress, Wikipedia), des vidéos (Youtube, Dailymotion), des images (Flickr, Picasa), etc. Les annotations produites par les utilisateurs sont utilisées pour décrire le contexte ou la sémantique de ces objets. Ils peuvent être divisés en :

  a. Structurés (des annotations sémantiques, également appelés concepts), où, semi-structurées (des balises sociales).
  b. Non structurées (par exemple texte libre, également appelé flux de sensibilisation sociale ou des mises à jour de statut).

La manipulation des objets implique des tâches telles que la description, la récupération, la réutilisation, la présentation et la recherche. Toutes ces tâches ont besoin d'une couche de connaissance préalable, qui est représenté par les annotations. Dans le cas de l'annotation automatique, le système extrait automatiquement les caractéristiques de l'objet (par exemple des descripteurs pertinents pour une image, des mots-clés à partir d'un document textuel, etc.) et les utilise comme annotations. Dans le cas de l'annotation semi-automatique, le système extrait généralement les annotations de la ressource, mais demande à l'utilisateur de les valider. Dans le cas des annotations manuelles, la capacité cognitive de l'utilisateur à interpréter le sens d'un objet est un effet de levier. Dans ce cas, l'utilisateur a généralement deux possibilités pour l'annotation : (i) le système donne à l'utilisateur une totale liberté dans le choix du terme qu'ils ont l'intention d'utiliser dans l'annotation (le cas de marquage social et texte libre annotations). Où, (ii) le système utilise un vocabulaire et l'utilisateur peut choisir un terme du vocabulaire pour l'annotation (le cas des annotations sémantiques et aussi quelques cas de l'étiquetage social). Cette deuxième option permet aux utilisateurs moins de liberté, mais permet d'avoir, un vocabulaire convergent stable qui permet une meilleure façon de récupérer des documents, que la description des ressources ne sera pas souffrir de synonymes, des erreurs d'orthographe ou de divergences dans la granularité. De plus, cette structure du vocabulaire de base peut encore être utilisée pour le calcul des similarités entre les ressources annotées. Dans ce qui suit, nous nous concentrons notre attention sur les annotations partagées dans les plates-formes sociales.

Les annotations peuvent être soit structurée, semi-structurée ou non structurée : Annotations structurés, dans ce cas, les termes employés dans l'annotation sont réglementés par un vocabulaire de domaine commun qui doit être utilisé par les membres du système. Ces types d'annotations ne sont actuellement pas utilisés dans la majorité des plates-formes sociales, car un vocabulaire de domaine contenant les termes nécessaires pour les annotations est nécessaire. Bien qu'une telle approche présente de nombreux avantages (par exemple, absence de synonymes, et absence de différences de prononciation), ce n'est pas la façon naturelle de décrire les ressources dans les plates-formes du Web 2.0. Comme le domaine n'est pas bien définie et, par conséquent, il est très difficile de construire ces vocabulaires et d'établir un accord pour chaque terme utilisé. Au même temps, les annotations sémantiques serrent lourdes à utiliser par les gens, car il exige un effort cognitif supplémentaire pour

sélectionner les concepts d'ontologies de domaine existants. En outre, les annotations sémantiques fonctionnent bien dans les systèmes où le domaine est bien défini (par exemple, un système de partage des connaissances sur les gènes humains Yeh, et al. (2003), mais dans les plates-formes sociales ce n'est pas le cas, car le contenu partagé est généralement très hétérogène, que les gens peuvent discuter sans limites (i.e. couvre plusieurs domaines sans régularités et relations).

Les annotations semi-structurés, en revanche, les annotations semi-structurées, telles que les balises sociales, sont largement utilisés dans les plates-formes sociales pour le marquage de photos et bookmarking (par exemple l'annotation d'une page Web). Ces annotations sont généralement choisies librement, ce sont des mots-clés sans un vocabulaire en arrière-plan. Cependant, nous les considérons comme semi-structuré, car ils représentent une approche intermédiaire entre les annotations sémantiques (des annotations qui sont basées sur des concepts d'ontologies de domaine et les annotations de texte libre). D'ailleurs, ces collections d'étiquettes convergent vers une organisation structurée de données, appelé folksonomie Gruber, (2005). Les annotations non structurées, une forme plus récente d'annotations, elle est représenté par des annotations libres de texte, également appelés flux de sensibilisation sociale, composée des mises à jour de statut ou microposts Naaman, et al., (2010). Cela peut être trouvé dans la plupart des réseaux sociaux et des systèmes de micro-blogging et se compose principalement de textes libres sous la forme de messages courts décrivant une ressource, une constatation, une impression, un sentiment, une activité récente, l'humeur ou futur plan. Les limites de cette pratique du point de vue de la recherche d'information et de gestion des connaissances sont similaires à celle de l'étiquetage social, car les utilisateurs ont toute liberté dans la formulation de ces messages. Il est important de mentionner que le contenu produit contient souvent lui-même la ressource décrite, sous la forme d'un lien hypertexte intégré. Une pratique courante est d'exprimer une opinion sur une ressource (par exemple page Web) ou de fournir son bref résumé à la communauté.

Depuis, Internet a repris Usenet comme le principal moyen de communication par ordinateur et il est passé par plusieurs étapes :

Le Web 1.0, un Web pour une lecture seule, avec de gros morceaux d'informations sous forme d'article de magazine de tailles importantes.

Le Web 2.0 : un Web de lecture-écriture, avec des forums imitant usenet, échanger des éléments d'information jusqu'à une demi-page dans la taille ; blogging, proche du

modèle de page Web, mais avec un décalage dans la paternité à pupilles du grand public ; et de micro-blogging, basé sur des messages très courts (140 caractères sur Twitter).

Ce passage de grande information, faisant autorité à l'information très court et amateur est contemporaine de l'évolution de la mobilité, avec les appareils plus conviviaux compatibles au Web (par exemple les Smartphones) mettant l'accent sur un facteur particulier : le contexte dans lequel l'information est écrite. Cela a brouillé la distinction entre l'information et la messagerie, que toutes les informations sur Twitter sont en fait un message aux disciples, et tous les messages peuvent être partagés, créant ainsi des informations.

1.2.1.2. Classification des plates-formes sociales selon le contenu partagé

En se basant sur le type de contenu partagé et sur le type de la ressource qu'il décrit, Stan, (2011) considèrent la classification suivante des plateformes sociales :

a. **Plates-formes de réseautage social –** (en anglais Social Network Platforms-SNP) : ces plateformes ont été l'origine d'une large acceptation des medias sociaux sur le Web. Ce genre de plateforme social a conduit la tendance vers le développement des données sociales. Ces plates-formes offrent la forme de base pour relier les pairs, et pour communiquer, commenter, partager, réagir, etc. des ressources et des personnes. Aujourd'hui, les réseaux sociaux sont devenus l'activité la plus populaire sur le Web.

b. **Plates-formes de publication sociale** - (en anglais Social Publishing Platforms-SPP) : elles partagent certains des mêmes fonctions que le premier type, mais avec un accent sur le contenu de l'édition. Cette catégorie de plates-formes est aussi appelée microblogging. Cela inclut les blogs, microblogs, ainsi que d'autres contenus partagés tels que les signets (Bookmarks) et les étiquettes (tags).

c. **Plates-formes d'agrégation sociale** - (en anglais Social Agreration Platforms-SAP) : ces plateformes existent pour l'objectif d'augmenter une ou plusieurs applications des réseaux sociaux les plus fréquent, généralement en fournissant une interface commune pour parcourir et / ou publier sur plusieurs sites de réseaux sociaux sous-jacents. Ces agrégateurs sociaux peuvent se différentier avec des fonctionnalités supplémentaires, l'ergonomie ou l'analyse des données sous-jacentes.

d. **Plates-formes de recherche social** - (en anglais Social Search Platforms-SSP) : elles sont définies comme les moteurs de recherche spécifiquement dédiés aux données sociales, y compris le contenu social (tels que les médias partagés et les commentaires)

et les relations sociales (telles que la recherche d'une personne dans un réseau avec certaines caractéristiques ou compétences). Cela signifie que ces outils indexent le contenu social et offrent un moyen aux utilisateurs afin de rechercher un contenu, d'une façon similaire à ce que fait Google avec le contenu des pages web en ligne.

e. **Plates-formes de marketing social -** (en anglais Social Marketing Platforms – SMP) : ce sont des applications qui exploitent les données sociales à des fins de marketing viral. Ce sont principalement des sociétés créées pour renforcer le modèle commercial des sites de réseautage social et de faire des recettes à partir des quantités énormes de données dont ils disposent.

f. **Plates-formes de géolocalisation sociale -** (en anglais Social Geolocation Platforms – SGP) : elles sont des plates-formes qui exploitent la mobilité des personnes et de leurs appareils, et leur permettent de déclarer leur présence et activités de leurs relations sociales. Ces plates-formes offrent les moyens de commenter différents objets associés à un emplacement physique bien défini et de s'engager dans une conversation avec des gens qui sont présents ou connectés à l'emplacement donné.

Le tableau 1.1 présente un panorama des plateformes sociales en ligne les plus populaires pour chaque catégorie mentionnée précédemment.

SNP	SPP	SAP	SSP	SMP	SGP
Facebook	Twitter	FriendBinder	Aardvark	Wildre	Aloqa
Myspace	Delicious	FriendFeed	Oneriot	Hypios	Foursquare
Secondlife	Twine	Silentale	Sysomos	Gravity	
LinkedIn	Yammer	Jodange	Swotti	Contexa	
Netlog	Posterous	Spokeo	Sproose	43 Things	
Hi5	Vox	Minggl	Moodviews		
Couchsurng	StumbleUpon	Yonoo	Eurekster		
Habbo					
Tribe					

Tableau 1.1 : Un panorama des plateformes sociales en ligne les plus populaires pour chaque catégorie Stan, (2011).

Cette classification est faite depuis 2011 ; aujourd'hui, avec les améliorations subies au sein des sites des médias sociaux, il existe de nombreux sites qui cernent toutes ces caractéristiques ou bien la plupart entre elles sous le toit d'une même plateforme. Comme un exemple concret, aujourd'hui, Facebook n'est plus qu'un SNP, il est aussi devenu une SSP, SMP et SGP.

1.2.1.3. Participation de l'utilisateur dans les plateformes sociales

Quelles sont les principales motivations des utilisateurs de partager du contenu dans une plate-forme sociale ? Cette question a déjà été étudiée par plusieurs travaux antérieurs, Dans cette section, nous essayons de fournir un bref résumé sur l'un d'entre eux. Le travail dans Naaman, et al., (2010) a décrit la motivation des utilisateurs afin de partager de contenu dans une plate-forme sociale :

- Le partage d'information : le partage des pages Web, principalement intéressantes avec la communauté.
- Autopromotion : le partage d'information sur des projets personnels, des blogs ou des compétences
- Opinions : le partage d'un avis positif ou négatif sur des produits, des technologies, etc.
- Les déclarations et les pensées aléatoires : le partage des sentiments ou des états actuels de l'esprit.
- Moi maintenant : le partage d'information sur l'activité actuelle.
- Question : le partager d'un besoin d'information avec la communauté.
- Présence actuelle : le partage d'informations sur l'emplacement actuel.
- Anecdote moi : le partage des histoires sur des événements, pourrait être intéressant de la communauté.
- Autres anecdote : le partage des histoires sur des événements curieux liés à d'autres choses.

Aujourd'hui, les réseaux sociaux sont devenus l'activité la plus populaire sur le Web.

1.2.2. Les réseaux sociaux

En sociologie, un réseau social est un concept qui date depuis les années 50. En 1954, ce concept a été mentionné pour la première fois par l'anthropologue britannique John A. Barnes.

Aujourd'hui, ce terme est sur toutes les lèvres, mais, quand cela a-t-il réellement débuté ?

1.2.2.1. Historique des réseaux sociaux

- ✓ USENET (1979) : inventé par un groupe d'étudiants de Caroline du Nord ; c'est un système de réseaux de forums qui permettait de publier et lire des articles, il est devenu par la suite un site d'échange de fichiers et de téléchargement.
- ✓ Le Minitel (1982) : c'est un produit 100% français, il est désigné aux services de renseignement téléphonique, services de messagerie (les premières expériences de messagerie instantanée), et autres services payants.
- ✓ Geocities, Caramail, et les premiers sites web (Année 1990) : après la naissance du World Wide Web en 1991, les premiers sites Web personnels ont vu le jour, grâce à la plate-forme

d'hébergement Geocities. Puis Caramail, un célèbre service de messagerie et site communautaires gratuit.

- ✓ Wikipedia et les premiers réseaux sociaux (années 2000) Wikipedia (2000)
- ✓ Friendster (2002), c'est le premier site Web fondé sur le concept de réseau d'amis (le tout premier réseau social). Rapidement dépassé par ses concurrents comme Myspace.
- ✓ Myspace (2003), il permet à ces membres de resauter et de disposer d'un espace personnel permettant de se présenter, diffuser ses photos et créer un blog.
- ✓ Wordpress (2003) : il permet la création des blogs personnels, c'est un logiciel libre et gratuit.
- ✓ LinkedIn (2003) : c'est le premier réseau social professionnel.
- ✓ Flicker (2004) : c'est un site de stockage et de partage de photos en ligne.
- ✓ Facebook (2004) : son nom d'origine Thefacebook, il a vu le jour dans une chambre d'étudiants de Harvard, avec Mark Zuckerberg à sa tête.
- ✓ Youtube (2004) : c'est un site de partage de vidéos.
- ✓ Twitter (2006) : il permet de publier des messages limités à 140 caractères.
- ✓ Foursquare (2007) : il est fondé sur la géolocalisation, il permet à ces utilisateurs d'indiquer qu'ils se trouvent dans un endroit précis, et de le recommander aux autres.
- ✓ Instagram (2010) : c'est la premieres application qui permet de partager facilement les photos prises avec un smart-phone sur internet.
- ✓ Pinterest (2010) : il est fondé sur le partage de contenus en relation avec les centres d'intérêt de ses utilisateurs.
- ✓ Google+ (2011) : faisant partie des produits Google. Il n'est pas radicalement différent de Facebook, mais ne retrouve pas du tout le même succès.
- ✓ Les années 2010 : le mobile social
- ✓ Spotify : il est soit gratuit avec publicité ou sur abonnement mensuel, il donne accès à de la musique en écoute illimitée.
- ✓ Snapchat : une application permettant d'échanger des photos qui seront instantanément effacées après avoir été visualisées.
- ✓ Tinder : une application de rencontre sur smartphone, fondée sur le concept de « match », ce match est obtenu lors d'une validation mutuelle pour permettre à deux membres d'entrer en contact.
- ✓ Vine : une application de partage de vidéos spécialement conçue pour les smartphones et les appareils nomades, ou ces vidéos durent six secondes et sont lues en boucle. Il appartient à Twitter.

1.2.2.2. Définitions des réseaux sociaux

Un réseau social peut être défini comme étant l'ensemble des interactions sociales qui unissent un groupe d'individus, ou chaque individu peut avoir plusieurs types de relation : amical, familial, professionnel ou entre regrouper pour un intérêt spécifique (sportif, culturel...) Bachelet, (2014). Depuis 2004, les réseaux sociaux sont déployés partout dans le monde. Ils ont été adoptées par les internautes avec une vitesse incroyable. Ils permettent aux utilisateurs d'établir des connexions ou des relations entre eux, comme l'amitié sur Facebook, ou la relation suivre sur Twitter.

1.2.2.3. Les catégories principales des réseaux sociaux selon leurs objectifs et usages

On distingue plusieurs types d'usage des réseaux sociaux :

- ✓ Les réseaux sociaux de partage de contenus : ils permettent aux utilisateurs de partager du contenu : vidéo, photo, sons, lien, document. Exemple, Youtube, Flickr, Vimeo, Myspace, Pinterest.
- ✓ Les réseaux sociaux destinés au partage d'expression : ils permettent d'évoquer des paragraphes sous forme d'articles, appelés « posts » : ce sont les blogs, les microblogs et les wikis. Exemples, Twitter, Google+, Tumblr, WordPress, Wikita.
- ✓ Les réseaux sociaux professionnels : leur objectif principal est de resauter, collaborer et rassembler une communauté d'intérêts professionnels par exemple au sein d'une même entreprise. Exemples, Viadéo, Linkedin, Glassdoor, Yammer.
- ✓ Les réseaux sociaux de géolocalisation : ils sont exclusivement liés à l'emploi d'appareils nomades, ils permettent aux utilisateurs de se localiser et partager leur position ou leur présence dans un lieu. Exemples, Foursquare, Swarm, Highlight, Glancee.
- ✓ Les réseaux sociaux destinés aux rencontres : ils promettent à ces utilisateurs de rencontrer des personnes (amitié, amour, etc.). Exemple, Badoo, Tinder, Happn, Once.
- ✓ Les réseaux sociaux de jeux : ils sont pour un objectif précis, qui est la fédération des communautés fans d'un ou de plusieurs jeux vidéo. Exemple, Twitch, Playfire.
- ✓ Les réseaux sociaux généralistes : ce sont les réseaux sociaux qui se trouvent à la croisée de plusieurs catégories des catégories cité précédemment. Remarquant que le phénomène social « Facebook », n'est pas cité en exemple des catégories précédentes, ceci est parce que le réseau social Facebook est un réseau généraliste qui permet de s'exprimer, de partager des contenus, de se localiser, de collaborer de façon professionnelle, et de jouer.

1.2.2.4. Les réseaux sociaux comme source de données

Les réseaux sociaux sont aujourd'hui des sites Web multi-usager, où nous pouvons communiquer avec des amis, lire les nouvelles, jouer à des jeux, rejoindre des groupes d'intérêt, les fichiers multimédias d'actions, et bien plus encore.

L'utilisateur d'un réseau social attribue des informations d'intérêt pour ses abonnés et essaye de rejoindre d'autres (notamment par la récupération). Au cours des dernières années, les réseaux sociaux ont connu une croissance explosive et a conduit à la génération d'énormes volumes de données liées à l'utilisateur. En fait, ces réseaux sociaux offrent une nouvelle façon de recueillir des données en quantités massives en temps réel. Ces données peuvent être analysées de différentes manières et ont donné naissance à un nouveau domaine de l'analyse de données, appelé Analyse des réseaux sociaux Kaplan and Haenlein, (2010). En visant l'analyse et la découverte de connaissances à partir des données sur ses utilisateurs et leurs intérêts. La plupart des travaux prévus dans la littérature de ce domaine sont adaptés en vue de résoudre des besoins spécifiques, qui comprennent, la découverte des tendances, l'enrichissement du contenu, le profil des utilisateurs, le regroupement basé sujet, l'analyse des sentiments, etc.

Wellman, (2001) argumente que les relations en ligne forment des réseaux sociaux virtuels représentatifs des réseaux sociaux réels. En effet ces réseaux virtuels sont créés à partir d'interactions initiées par des personnes physiques. Cet argument est confirmé par Mika, (2007) mais il souligne le caractère incomplet de ces réseaux sociaux en raison de l'absence en ligne de certaines composantes de la réalité.

1.2.3. Les caractéristiques du contenu social

Les sites des réseaux sociaux encouragent leurs utilisateurs de s'exprimer et de partager des tons de données. Cette interactivité amène un nouveau type de texte, qui présente des caractéristiques spécifiques, ce texte est libre, massive, et dynamique. Habituellement, les internautes ne se soucient pas de l'orthographe ni de la construction grammaticale exacte d'une phrase. Ils parlent avec courage en utilisant les termes d'argot (acronymes et abréviations) dans leurs messages. Parfois, les utilisateurs utilisent plusieurs unités lexicales qui semblent syntaxiquement différentes mais décrivent en fait le même sens. Par exemple, («$», «dlrs» ou «dollars») représentent la même chose, «dollars», («2 $», «50 $», «756 $», «1dollars») représentent «une somme d'argent». ", (" $ "," Dollars ", "£", "¥") représentent "une monnaie" ou, en général, tous ceux qui représentent "de l'argent" (Benkhelifa, and Laallam, 2016).

En outre, un post est caractérisé par sa forme libre dans les plates-formes sociales, ainsi que sa

taille courte. Ce dernier est émergé comme un moyen simple et pratique pour interagir et communiquer d'une manière très facile et libre. La taille de ces posts est limité par quelques plates-formes social, l'exemple le plus courant et connu est Twitter, où les tweets sont limités par 140 caractères.

De nouveaux moyens ont émergés afin de mieux identifier les messages pertinents à un événement spécifique, (appelé hashtags). D'autres moyens sont émergés pour synthétiser une information, tels que l'élimination des Stopwords afin de garder que les termes essentiel (mots-clés et les entités nommées). Ces pratiques dépendent en grande partie de la communauté des utilisateurs ciblés, qui peut varier dans le monde entier. Aussi compte tenu de sa taille courte, ces microposts sont souvent appelés signaux sociaux Mendes, et al., (2010), et les utilisateurs de ces systèmes capteurs sociaux Sakaki et al., (2010), car ils peuvent être utiles pour détecter des événements importants dans un endroit donné, comme un tremblement de terre.

1.2.4. Les médias sociaux

Un autre terme qui est extrêmement cohérant dans ces dernières années, « les médias sociaux » : c'est une notion plus générale qui englobe les réseaux sociaux et désigne généralement l'ensemble des sites et plateformes sociaux du Web.

Dans cette thèse on s'intéresse spécialement aux réseaux sociaux et non à tous les médias sociaux, qui ne portent pas la notion d'un réseau, au sens strict du terme, par exemple, les forums et les blogs sont des médias sociaux mais ne sont pas des réseaux sociaux.

Tous les jours, les plateformes sociales et les réseaux sociaux génèrent des quantités confédérales des données personnelles et sociales. Ces données représentent une source d'information précieuse pour plusieurs applications. Malheureusement, l'accès à ces données devient de plus en plus difficile. Cependant, organiser, modéliser et accéder rapidement à l'information devient alors un réel challenge face à cette grande masse de données.

Plusieurs études portent sur la modélisation des caractéristiques des utilisateurs et des interactions dans les réseaux sociaux. La partie suivante montre les efforts déployés par la communauté du Web sémantique pour résoudre ce problème.

1.3. Le Web Social Sémantique

1.3.1. Introduction

Le Web social a été largement adopté et ce, depuis sa naissance. Il permet l'interaction et la

participation sociales à travers la création des espaces sociaux tels que les blogs, et les réseaux sociaux. Malgré la nature dynamique et profondément interconnecté des ressources du Web social en représentant un genre particulier de données, elles sont limitées en termes de réutilisation de ces données sociales. Ces derniers peuvent ajouter une valeur précieuse si elles sont facilement connectables et accessibles. Des technologies et des normes clés du Web sémantique évoluent parallèlement au Web social. Les espaces sociaux du Web social peuvent être combinés aux technologies du Web sémantique pour catalyser la prochaine étape du Web et permettre la création de nouvelles applications en termes de la découverte de connaissances et l'exploration de données. Nous allons décrire cette convergence dans ce chapitre.

1.3.2. Le web Sémantique

Le web sémantique a été défini par Berners Lee : ≪ Le Web Sémantique est une extension du Web actuel dans laquelle l'information est enrichie de sens bien défini, permettant aux humains et aux ordinateurs de mieux travailler ensemble. ≫ Berners et al., (2001).

En particulier, le Web sémantique utilise des identificateurs de ressources uniformes (URI) afin d'identifier les entités disponibles sur le Web sous une forme de documents ou d'images, mais ne pas en tant que concepts ou propriétés. Le cadre de la description des ressources (RDF : ressources description framework) (http://w3.org/ TR / REC-rdf-syntaxe) décrit ces ressources dans un format interprétable, dont le bloc de construction de base est une déclaration triplets objet-attribut-valeur. Les ressources peuvent être (des auteurs, des livres, des lieux, des personnes, des hôtels, etc.), tandis que les propriétés décrivent les relations entre les ressources telles que « écritPar », « âge », « titre », etc.

Pour fournir des représentations traitable et accessible par la machine, il est habituel de coder les triplets RDF en utilisant une syntaxe XML. Cependant, RDF ne fait pas d'hypothèses sur tout domaine d'application particulier, ni définit la sémantique de tout domaine. A cet effet, il est nécessaire d'introduire des ontologies qui permettent de bien traiter et représenter des connaissances.

Les ontologies

Selon la littérature du Web sémantique, l'ontologie est définie comme une spécification explicite de la conceptualisation d'un domaine Gruber, (1993). Les ontologies permettent le partage de la compréhension commune sur un sujet, soit, entre des personnes ou des agents logiciels. En outre, les ontologies permettent le raisonnement, à partir des données et des informations supplémentaires exprimées sous la forme d'une ontologie, et de déduire de nouvelles relations entre les données. Différents langages ont été développés pour la conception des ontologies, dont,

les plus populaires sont, RDFS (RDF Schema) (http://w3.org/TR/rdf-schema) et OWL (Ontology Web Language) (http://w3.org/TR/owl-fonctions).

Une référence importante dans la gestion des métadonnées sémantiques spécifiquement dans le cas des réseaux sociaux est Mika, (2007), où l'auteur enquête sur le bénéfice mutuel des réseaux sociaux et le web sémantique et plus spécifiquement comment les deux peuvent être combinés ensemble.

De nos jours, de plus en plus de données sont disponibles sur le Web 2.0. De plus en plus de personnes et d'organismes publics/privés contribuent à ce déluge en choisissant de partager leurs données avec les autres. Aujourd'hui, de nombreuses applications dans différents domaines consomment ces données sociales pour créer de nouvelles opportunités d'affaires et d'améliorer les stratégies de recommandation, tels que le marketing en ligne. De toute évidence, l'évolution future de cet écosystème nécessite la présence de données structurées, ce qui soulève certainement la question de la façon de fournir l'accès à ces données afin qu'elles puissent être réutilisable par la suite. La communauté du Web sémantique est basée sur le principe que les données publiées dans le Web et plus spécialement le Web social devraient prendre la forme la plus structurée, et la plus facile à retrouvées. Cette intersection du Web sémantique et du Web social est appelée le "Social Semantic Web" en français « le Web social sémantique ».

1.3.3. Le Web social sémantique

Le Web sémantique social vise à résoudre certaines des limitations susmentionnées auparavant, tout en combinant la structure du Web social et les normes du Web sémantique, en créant une plate-forme permettant d'avoir des espaces sociaux sémantiquement améliorés.

Le Web social sémantique peut être vu comme une plate-forme d'échange collaboratif personnel et professionnel avec des contributions communautaires réutilisables Omitola et al. (2014) Grâce à l'utilisation des données du Web sémantique, un contenu interrogeable et interprétable est ajouté aux infrastructures de collaboration existantes du Web social, et une utilisation intelligente de ce contenu peut être réalisée à l'intérieur de ces espaces sociaux sémantiquement améliorés - permettant la vision de données sémantiques à Web à réaliser au maximum de ses avantages.

Comme nous l'avons dit précédemment, la structure actuelle des plateformes sociales ne permet pas d'extraire de manière rapide et simple les informations cachées recherchées par les applications dans le monde réel. Le défi consiste donc à restructurer et à modéliser ces données de manière à faciliter leur accès et leur exploitation ultérieure par ces applications. Plusieurs études portent sur la modélisation des caractéristiques des utilisateurs et des interactions dans les

réseaux sociaux. Des efforts importants ont été déployés par la communauté du Web sémantique pour résoudre ce problème en essayant de représenter l'utilisateur le plus précisément possible. Les vocabulaires représentés dans un format standard, tels que RDF (Resource Description Framework) et OWL (Web Ontology Language), ont été développés pour modéliser sémantiquement les utilisateurs et leur contexte social.

1.3.4. Les vocabulaires du Web social sémantique

1.3.4.1. Ontologies pour modéliser les profiles utilisateurs

La modélisation/le profilage de l'utilisateur, en termes généraux, peut faire référence à différentes significations sans définition spécifique. Une définition générale du profilage de l'utilisateur donnée par Zhau, et al., (2012) est « le processus consistant à acquérir, extraire et représenter les caractéristiques des utilisateurs ».

Des efforts importants ont été déployés par la communauté du Web sémantique pour développer des vocabulaires permettant de modéliser les caractéristiques des utilisateurs et des interactions dans les réseaux sociaux. Nous discutons ci-après les plus importants et les plus connus :

- ✓ **FOAF – Friend-of-a-Friend**

  Le modèle sémantique le plus connu pour les profils utilisateur est certainement Friend-of-a-Friend (FOAF). Un vocabulaire RDF ciblé pour décrire les profils, les amis, des liens vers des contacts, des services, etc. FOAF entend fournir un moyen de représentation des informations sur les gens d'une manière simple et facile à traiter. L'idée derrière FOAF est celle d'un réseau social lisible par une machine complètement décentralisée qui repose sur des profils personnels. Du point de vue de l'utilisation, une personne est décrite en utilisant la classe foaf:Person, énumérant les attributs clés tels que le nom, le sexe et les ressources connexes. En pouvant également faire la liste de leurs intérêts. Chaque personne est identifiée de manière unique en utilisant la propriété foaf:mbox contenant son adresse e-mail. Aussi FOAF peut constituer le réseau d'amitié d'une personne. Où, chaque ami est une instance de la classe foaf:Person. FOAF est le profilage la plus utilisée pour exprimer des informations personnelles et des relations au sein des communautés en ligne.

- ✓ **Microformat**

  Microformat fournit des vocabulaires pour décrire les personnes, ainsi que leurs liens sociaux. Le microformat hCard (microformat hCard) représente les personnes et leurs attributs tels que le nom donné, nom de famille, URL, email, etc. Le microformat XFN (XFN) capture de réseaux sociaux en représentant les relations entre les personnes par l'intermédiaire de l'attribut 'rel' (par exemple, <a href="http://jeff.example.org" rel="friend

met">).

- ✓ **Schema.org**

  Les vocabulaires schema.org, (schema) ont accord avec les principaux fournisseurs de moteurs de recherche (Google, Bing, Yahoo! et Yandex), ils sont capables de capturer les connaissances sur les gens et leurs réseaux sociaux. Ils fournissent une collection de balises pour définir les types d'éléments (personne, Place, Organisation, avis, événement, etc.) et les relations sociales (sait, collègue, enfants, parents, frères et sœurs, associés à. Etc.).

- ✓ **SIOC**

  Le SIOC (semantically Interlinked Online Communauties) ontologie (SIOC), à l'origine conçu pour capturer la connaissance des forums de discussion. Il modélise non seulement les utilisateurs et les interactions sociales, comme dans les travaux mentionnés précédemment, mais aussi le contenu. Cette ontologie réutilise les classes et les relations de plusieurs ontologies bien connues telles que FOAF vocabulaire (FOAF) et Dublin Core Metadata Terms (dcterms).

- ✓ **OPO Model (Online Presence Ontology)**

  Le modèle OPO est une ontologie qui vise à modéliser les aspects dynamiques de la présence d'un utilisateur en ligne et pour permettre l'échange des données de présence en ligne. OPO est mentionné dans le rapport final du W3C Social Groupe Incubateur comme l'un des vocabulaires qui prennent en charge la publication de données sur le Web social sur le Web sémantique, ainsi que FOAF et SIOC.

1.3.4.2. Les modèles sémantiques pour capturer le contenu partagé dans les RS

Le contenu est certainement la partie qui a attiré plus d'attention dans le domaine du web social. Actuellement, de nombreux modèles sémantiques ont été conçus pour le contenu social. Les premiers modèles de Gruber et Newman., (2005) Tag Ontology sont à la base de la plupart des modèles de gestion de contenu. Ces modèles ne sont pas intégrés dans les applications web, mais sont toujours utiles car ils permettent de comprendre les principales idées derrière la gestion de contenu sémantique et ses dimensions principales.

L'objectif de ces modèles est de modéliser les informations relatives à l'utilisateur afin de construire une image réaliste au sujet de son contexte en les rendant portable d'une plateforme à une autre.

Un premier effort pour conceptualiser l'activité de marquage des ressources sur le Web est celui de Gruber, (2005). Une ontologie de représentation générale pour le tagging. Elle tente de créer une représentation commune et une infrastructure pour relier les taggers, les tags et les ressources. Ceci est une formalisation d'une activité sociale, à savoir celle de tagging du contenu

avec une ou plusieurs tags, choisissant librement des mots-clés.

- ✓ Newman et al., (2005) ont défini une ontologie de tags et tagging, appelé Tag Ontology, qui décrit la relation entre un agent (tagger), une ressource arbitraire, et un ou plusieurs mots clés.
- ✓ Meaning Of A Tag (MOAT) Passant, A., & Laublet, (2008) tente de fournir un modèle de Web sémantique pour définir la signification des étiquettes d'une manière lisible par machine. Pour y parvenir, MOAT définit les significations globales d'un tag, à savoir la liste de toutes les significations et aussi la signification locale d'une étiquette, à savoir la signification d'un tag dans une action de marquage particulier. Par exemple, le tag Paris peut signifier- en fonction de l'utilisateur, le contexte et d'autres facteurs - une ville en France.
- ✓ The Social Semantic Cloud of Tags (SCOT) Kim et al., (2008) a pour le but de décrire la structure et la sémantique des données de marquage et d'offrir l'interopérabilité sociale des données entre les sources hétérogènes.
- ✓ Simple Knowledge Organization System (SKOS) (Miles and Bechhofer,, 2009) est un vocabulaire RDF pour représenter semi-formels Knowledge Systems Organisation (Koss), tels que thesaurus, taxonomies, les systèmes de classification, etc.
- ✓ Communautés sémantiquement-Interlinked en ligne : SIOC Breslin et al., (2005) est un vocabulaire qui vise à interconnecter les sites en ligne communautaires et de discussions sur Internet. Bien que ce vocabulaire ne soit pas une Ontologie de tag, mais, ce modèle est toujours très important, car il donne un modèle plus général de la gestion de contenu du Web social. L'idée est de permettre aux multiplateformes de telle sorte que la conversation couvrant plus de multiples médias (par exemple, blogs, forums, listes de diffusion, etc.) en ligne peuvent être unifiés en un format ouvert. Le format d'échange exprime l'information contenue explicitement et implicitement dans les méthodes de discussion sur Internet, d'une manière lisible par machine. Communautés en ligne permettent aux utilisateurs Web d'exprimer leurs pensées, obtenir des commentaires et d'interagir avec des personnes qui partagent un intérêt similaire. Les utilisateurs du Web modernes ont tous une sorte de participation dans ce domaine : forums, salons, forums et sites de réseautage social de chat. Chaque communauté peut être considérée comme un jardin clos, sans lien à d'autres. Le projet SIOC se concentre sur les moyens d'intégrer et fusionner ces jardins clos, en fournissant des ponts entre les utilisateurs et les ressources qui existent dans chacun d'eux.
- ✓ Autres propositions pour capturer le contenu social sont les suivantes : LODR (A Open Data Système Tagging ing Link-) Passant, (2008), qui est très similaire à Moat, GroupMe! Ontologie de Abel, (2008) ajoute la contextuelle supplémentaire en formation à l'activité de

marquage (par exemple le temps, l'humeur de l'utilisateur, les connaissances de base de l'utilisateur, etc.) et le modèle conçu par Echarte et al., (2007) considère une dimension supplémentaire, la polarité de sentiment de tags.

Dans cette partie, nous avons introduit les concepts du Web sémantique et du Web social sémantique, les différents vocabulaires dédiés à la représentation et la modélisation sémantique des utilisateurs, des communautés en ligne et des objets sociaux existant dans le web social, en particulièrement dans les réseaux sociaux.

Les données sociales décrites sous forme RDF fournissant une représentation plus puissante et plus riche des réseaux sociaux par rapport à l’analyse classique des graphes des réseaux sociaux. Bien que ces ontologies modélisent les caractéristiques et les interactions des utilisateurs au sein des communautés en ligne, elles sont incapables de donner un sens au contenu, de modéliser des d'aspects plus dynamiques tels que les intérêts des utilisateurs, de regrouper les utilisateurs discutant de sujets, ou capturer les mêmes messages traitant du même sujet. Cependant, il faut approfondir l'exploitation du contenu (enrichir le profil d'utilisateur), ce qui dépasse les métadonnées structurelles créées à l'aide de ces ontologies. La partie suivante discute les techniques traitant ce problème.

1.4. Prétraitement et catégorisation du texte social

L'information est devenue un élément essentiel et indispensable pour tous les secteurs, et dans toute entreprise. Aujourd'hui, le défi n'est pas de chercher l'information, car elle est disponible avec des quantités gigantesques, mais de trouver l'information utile et pertinente afin de pouvoir prendre des décisions.

Dans cette partie, nous discutons les différentes techniques permettant d'avoir ces informations utiles et pertinentes. Nous commençons par donner des généralités sur le data mining, le text mining, l'apprentissage automatique, et le traitement du langage naturel. Par la suite, nous décrivant la tâche globale de la catégorisation de textes, tout en exposant les méthodes et quelques travaux réalisés dans la littérature pertinente. Ensuite, un apercevant les travaux réalisés relatives au traitement et la classification des textes issus des médias sociaux, avec un certain accent sur ceux qui sont employés pour la classification de texte issus des réseaux sociaux.

1.4.1. Le Data Mining

L'information est un élément clé dans de nombreux contextes. Où, toute une partie de la recherche scientifique s'intéresse à la collecte et l'analyse de données. Dans plusieurs contextes, les informations sont recueillies et utilisées pour prendre des décisions.

Néanmoins, deux majeurs problèmes s'imposent au sein de cette vision, le premier problème est le manque d'information. Généralement, ceci signifie que les données possédées ne sont pas assez pour avoir suffisamment de connaissances précises sur le phénomène qui est en cours d'analyse. Le deuxième problème est l'excès d'informations (beaucoup trop d'informations) due au développement continu des outils de stockage, de traitement, et de communication. Ce qui rend l'extraction de l'information utile à partir d'une quantité trop abondante de données une tâche très difficile, fastidieuse et presque impossible pour être faite par des humais.

À cette fin, les deux disciplines le Data Mining et le Knowledge discovery ont vu le jour, ces disciplines s'intéressent au développement des méthodes et des techniques pour l'extraction non-trivial de l'information de haut niveau de quantités potentiellement énormes de données brutes Frawley et al., (1992). Le Data Mining permet de construire des modèles prédictifs, et de partitionner les grands ensembles en plus petits, en construisant des groupes en se basant sur la similarité des caractéristiques entre les éléments au sein du même groupe. Il est principalement destiné à l'analyse des données structurées et appliqué avec succès dans de nombreux domaines. Compte tenu du grand besoin du data mining dans une large variété de domaines qui traitent des données brutes et non structurées telles que le texte, plusieurs outils et techniques ont été

développée. Par exemple, due aux certains facteurs, tel que la croissance de données sur les utilisateurs et leur comportement en ligne, de plus en plus d'activités sociales en ligne ; le coût décroissant de la collecte, du stockage et du traitement des données ; et l'expansion des plateformes de médias sociaux. Il est nécessaire d'appliquer le data mining sur les données issues des médias sociaux afin d'analyser, organiser, classer et donner du sens à ces données, en utilisant des techniques de l'analyse de réseaux sociaux, le text mining, le machine learning et le traitement du langage naturel.

1.4.2. L'apprentissage automatique

Le processus du Data Mining se base sur de différentes étapes, allant de la collecte et le traitement des données à l'analyse des résultats finaux. L'étape clé est évidemment la conversion de la grande quantité de données brutes, à des informations de haut niveau qui peuvent être facilement interprétés. Cette transition est généralement basée sur des techniques d'apprentissage automatique (machine learning). Ces techniques sont généralement utilisées pour construire des modèles de connaissances à partir des données. Ces modèles peuvent mettre en évidence des caractéristiques intéressantes à travers les données analysées afin de les utilisés pour faire des prédictions concernant les données arrivant à l'avenir.

Définitions

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle. Il a été défini en 1959 par Arthur Samuel, un pionnier dans le domaine de l'intelligence artificielle et de l'apprentissage automatique qui a fait le premier programme d'auto-apprentissage. Samuel a défini l'apprentissage automatique comme *« un domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmé »*. L'apprentissage automatique est basé sur la construction de programmes informatiques qui s'améliorent automatiquement avec l'expérience Mitchell, (1997). Il repose sur la mise en œuvre d'un algorithme, ou, plusieurs algorithmes d'apprentissage automatique existent, ces derniers peuvent s'appliquer selon le type de la tâche qu'ils réalisent.

1.4.2.1. Type d'apprentissage

Les méthodes d'apprentissage automatique peuvent être classées en : supervisées, non supervisées :

- ✓ **L'apprentissage supervisé**

  Le but principal du processus de la classification supervisé est de regrouper des éléments qui ont des caractéristiques communes en catégories distinctes. « L'apprentissage supervisé

consiste à établir des règles de comportement à partir d'une base de données contenant des exemples de cas déjà étiquetés » Laloë, (2009).

En général, ce processus implique la construction d'un ensemble de données correctement marqué avec des labels d'un ensemble de classes bien défini, et un algorithme d'apprentissage qui analyse ces données pour trouver des corrélations pertinentes entre les caractéristiques de ces données et ses labels afin de construire un classifier, ce dernier doit être en mesure d'étiqueter correctement toute donnée représentée avec les mêmes caractéristiques.

La classification supervisée est l'une des tâches les plus fréquemment exécutées par les systèmes dits intelligents Kotsiantis et al., (2007). Ainsi, un grand nombre d'algorithmes ont été développées. Dans cette section nous nous concentrerons sur les algorithmes les plus importantes.

1- Algorithmes basé sur la logique

   a- Arbre de décision

   b- Ensemble de règles d'apprentissage

2- Techniques à base de perceptron

   a- Perceptrons monocouches

   b- Perceptrons multicouches

   c- Réseaux de fonction de base radiale (RBF)

3- Algorithmes d'apprentissage statistique

- ✓ **L'apprentissage non supervisé (le clustering)**

Le clustering est la tâche de partitionnement des données en les mettant dans certains groupes (appelé : clusters), où aucune connaissance préalable n'est donnée. L'objectif d'un algorithme de clustering est : étant donné un ensemble de données, dont l'objectif est d'identifier les clusters dans cet ensemble de telle sorte ou chaque cluster contient des éléments qui ont des caractéristiques très semblables, et significativement différents de ceux des autres clusters. Également, de nouveaux éléments peuvent alors être affectés aux mêmes groupes. Contrairement à la classification, les algorithmes de clustering sont non supervisés.

1.4.2.2. Le deep learning

« Le deep learning est un nouveau domaine de recherche de l'apprentissage automatique. Il consiste à apprendre plusieurs niveaux de représentation et d'abstraction qui aident à

donner un sens aux données tels que des images, du son et du texte. » Moualek (2017).

Généralement, avant tout apprentissage machine, les données d'entrée aux algorithmes doivent subir une très importante étape de prétraitement. Les méthodes d'apprentissage sont généralement en mesure de travailler uniquement avec des types spécifiques de données : certains d'entre eux dans leurs versions de base ne peut faire face à des valeurs discrètes, telles que des nombres entiers ou des données catégoriques (éléments à partir d'un ensemble fini), tandis que d'autres sont spécialisés sur des gammes continues de valeurs, qui peuvent de toute façon être discrétisées afin d'être traités par d'autres algorithmes.

Les algorithmes d'apprentissage automatique, tel que discuté ci-dessus, sont aptes à gérer des données structurées. Un type de données non structuré comme le texte brut et libre écrit en langage naturel ne peut pas être traité directement par ces algorithmes.

1.4.3. Le Text Mining

Dû à l'augmentation du nombre des documents électroniques, et du web en particulier. Une grande importance a été mise sur la fouille de texte (text mining). Ces documents comprennent souvent des données non structurées ou par fois semi structurées. Ceci représente le défi principal du text mining. Le text mining est défini comme l'extraction des informations à partir des ressources textuelles à travers l'application des opérations telles que, la classification (supervisée et non supervisée) et la synthèse. Text mining est le domaine de la recherche qui porte sur l'extraction de l'information utile et de haut niveau du texte. Pratiquement, le text mining est le data mining appliquée sur des données textuelles écrites en langage naturel (Pasolini, 2015). En text mining, plusieurs domaines comme le Traitement du Langage Naturel (TAL), et l'apprentissage automatique travaillent ensemble pour modéliser les différents types de documents afin de les classer automatiquement.

Comme dans le DM, un processus du Text Mining implique généralement une phase préliminaire de prétraitement, nécessaire pour transformer le texte brut en une forme structurée (représentant les documents en vecteurs), afin que les algorithmes d'apprentissage automatique puissent être appliqués.

1.4.4. La catégorisation de texte

La classification du texte est une partie importante du text Mining, la catégorisation de texte (également connu sous le nom de classification de texte) est la tâche de trier automatiquement un ensemble de documents dans un ensemble prédéfini de catégories (ou classes) Sebastiani, (2005). Ces documents peuvent être classés en utilisant les méthodes d'apprentissage automatique tel que

(naïve bayes, decision tree, K-nearest Neighbors (KNN), Support Vector Machines (SVM), et les réseaux de neuronaux).

Définition du problème Sebastiani, (2002) : La tâche générale de catégorisation de texte peut être formellement définie comme la tâche d'approximation d'une catégorie inconnue par une fonction d'affectation $F : D \times C \rightarrow \{0, 1\}$, où D est l'ensemble de tous les documents possibles $\{d_1 \ldots d_n\}$ et C est l'ensemble des catégories prédéfinies $\{c_1 \ldots c_m\}$. La valeur de $F(d, c)$ est 1 si le document d appartient à la catégorie c et 0 sinon. La fonction d'approximation $M : D \times C \rightarrow \{0, 1\}$ est appelé un classifier, et la tâche est de construire un classifier en utilisant des algorithmes d'apprentissage automatique pour faire des prédictions précises fondées sur des observations passées avec la production de résultats comme "proche" que possible de la véritable fonction catégorie d'affectation F.

La classification du texte a été appliquée à une grande variété d'applications pratiques :

- Comprendre et analyser les sentiments de l'audience à partir des médias sociaux (anamasr)
- Filtrage des emails spams et non-spam
- Marquage automatique des requêtes des clients
- Catégorisation des articles d'actualité topics en sujets prédéfinis catalogage des articles de presse

1.4.5. Traitement du langage naturel

Les gens communiquent en langage naturel. Le domaine de traitement automatique des langues (TAL) (en anglais NLP : Natural Language Processing) se réfère généralement à des méthodes et des techniques permettant à un ordinateur de gérer directement du texte en langage humain afin d'en extraire des informations utiles.

Afin d'obtenir des informations plus détaillées sur une phrase donnée, cette dernière doit subir à une analyse de différentes étapes, qui sont généralement suivantes :

- ✓ La morphologie : cette phase soucie des mots qui composent la phrase, de la manière dont ils sont formés et de la façon dont ils changent en fonction de leur contexte. Voici quelques exemples : (Préfixes / suffixes, Singularisation / pluralisation, Détection de genre, Lemmatisation (la forme de base du mot, ou l'inverse de l'inflexion), Vérification orthographique, etc.)

- ✓ La syntaxe (parsing) : cette étape concentre sur la relation des mots dans une phrase (comment une phrase est construite). L'analyse syntaxique est généralement effectuée au niveau de la phrase. Dans un sens, la syntaxe est ce que nous appelons habituellement la grammaire.

- ✓ La sémantique : ici, le sens de la phrase est analysé (par exemple, comprendre le sujet d'une discussion Twitter). Voici quelques exemples de tâches effectuées à ce stade : (la reconnaissance d'entité nommée (en anglais : Named Entity Recognition) et l'extraction de relation (en anglais : Relationship Extraction)).

- ✓ La pragmatique : cette étape essaye de comprendre le texte dans son ensemble. Elle essaye de résoudre des problèmes populaires comme (la modélisation de sujet, coréférence / anaphore, la récapitulation, et question/réponse).

1.4.5.1. La représentation de texte

Le principal obstacle rencontré lors du traitement automatique de texte est sa forme non structurée. Un ordinateur voit un texte comme une séquence de caractères, généralement sans aucune information sur les mots, les phrases ou les sections. Les structures grammaticales qui sont généralement facile à reconnaitre par les lecteurs humains ne peuvent pas être trivialement comprises par les ordinateurs, en particulier en raison des ambiguïtés sémantiques et syntaxiques. Pour cela il est nécessaire d'utiliser une technique de représentation efficace permettant de représenter les textes sous une forme exploitable par la machine.

Vue l'importance de la représentions des textes dans le processus de classification de texte. De nombreuses techniques récurrentes se sont avérées efficaces dans de nombreuses situations tout au long de la littérature. Cette session donne un aperçu de ces techniques, avec un certain accent sur ceux qui sont employés pour la classification de texte.

- ✓ **La représentation « Sac-de-mots » Bag of words**

  Une représentation en sac de mots devient l'une des méthodes les plus populaires pour représenter le contenu textuel et elle a été appliquée avec succès au texte catégorisation zhang et al., (2010). Ici, le mot est l'élément de base constituant un document texte. Où, chaque mot dans le corpus d'apprentissage étant traité indépendamment du mot précédent et du mot suivant Lundborg, (2017). En pratique, un sac de mots est un ensemble qui peut contenir plusieurs occurrences des mots contenus dans un document texte, sans tenir compte de leur ordre et à l'exclusion d'autres éléments comme des signes de ponctuation. Les textes sont transformés simplement en vecteurs dont chaque caractéristique correspond à un seul mot, selon le modèle d'espace vectoriel (Vector Space Machine VSM).

Bien que cette représentation implique une perte d'information significative, il se révèle être efficace pour accomplir de nombreuses tâches pratiques avec une bonne précision. Un exemple de classification de texte, le sujet traité dans un document peut être parfaitement indiqué en se basant sur les mots les plus récurrents dans ce document, étant donné que les mots représentatifs des catégories d'intérêt sont bien connus.

- ✓ **La représentation N-gram**
  - **Le modèle n-grammes de mots**
    Le modèle N-gramme prend en compte l'ordre des mots en concédant les séquences des mots, N représente le nombre de mots à inclure dans une séquence. Ce qui donne un sens plus profond aux phrases et aux textes.
  - **Le modèle n-grammes de caractères**
    Les n-grammes de caractères sont similaires aux n-grammes de mots, mais au lieu de créer des représentations vectorielles de mots et de combinaisons de mots, une représentation vectorielle est basée sur des caractères et des combinaisons de caractères.
- ✓ **Word2Vec**
  Word2vec est une technique de plongement de mots (en anglais : word embidding) introduite par Mikolov et al., (2013) ; Cette technique focalise sur l'apprentissage d'une représentation de mots. Ou chaque mot est représenté par un vecteur. La méthode word2vec est capable de calculer les degrés de similarité entre les mots. Elle utilise des contextes de mots pour modéliser la signification sémantique d'un mot lors de la fusion de synonymes.

Après avoir donné quelques représentations textuelles des documents, il est nécessaire d'adopter une stratégie de représentation de texte qui attribuer une valeur appropriée à chaque terme. Autrement dit, la réduction des documents textuels à des vecteurs dans l'espace des termes afin de pouvoir être traité par la machine. Ces stratégies sont appelées (Term weighting scheme).

1.4.5.2. Term weighting scheme (TWS)

La pondération des termes est connue sous le nom de stratégie de présentation de texte pour attribuer une valeur appropriée à chaque terme afin d'améliorer les performances de la classification de texte lors de la transformation du contenu d'un document textuel en vecteur dans l'espace des termes Feng et al., (2018).

Un schéma de pondération est utilisé pour convertir des documents de texte en vecteurs dans les espaces de terme, cette étape est essentielle dans le processus de la catégorisation automatique du

texte Lan et al. (2005).

Dans cette section, de différents schémas de pondération des termes supervisés et non supervisés sont monté. Les notations *a, b, c* et d sont utilisé pour désigner le nombre de documents diffèrent comme ci-dessous :

- ✓ a est le nombre de documents de la catégorie positive contenant ce terme
- ✓ b est le nombre de documents de la catégorie positive ne contenant pas ce terme
- ✓ c est le nombre de documents de la catégorie négative contenant ce terme
- ✓ d est le nombre de documents de la catégorie négative qui ne contiennent pas ce terme

N représente le nombre de tous les documents. Où, le tf représente la fréquence de terme brute (nombre de fois qu'un terme apparaît dans un document). Le tableau suivant représente quelques schémas de pondération utilisés dans la classification de texte, ainsi que leurs formations mathématiques.

Term weighting	Nom	Formulation mathématique
Binary	Binary	(1 pour la présence, 0 pour l'absence)
Term frequency - Inverse term frequency	TFIDF	$tf.log\frac{N}{a+c}$
Chi-square	ChiS	$tf.\frac{N(a*d-b*c)^2}{(a+c)(b+d)(a+b)(c+d)}$
Odds ratio	OddsR	$tf.log\frac{a*d}{b*c}$

Tableau 1.2 : Des différents schémas de pondération

1.4.6. La classification des textes issus des réseaux sociaux

Récemment, plusieurs recherches sur la classification de texte ont mis l'accent sur les messages partagés dans les médias sociaux. La majorité des travaux antérieurs s'intéressent à classer un type spécifique de messages partagés dans les médias sociaux dans des catégories spécifiques (par exemple la classification des news Noci et al., (2012), Weber, (2014), la classification des sentiments Liu, (2012), Akaichi, et al., (2013), Al-Ayyoub et al., (2015), Kermanidis, and Maragoudakis, (2013), Feng et al., (2019); Al-Smad et al., (2018) Zhang et al., (2018), Benkhelifa and Laallam, (2018), mais ils ne se soucient pas sur le sujet d'un poste aléatoire.

Dans les recherches Vanetti et al. (2013), Uttarwar, and Bhute (2013), Contratres et al., (2018) les auteurs fournissent aux sites des réseaux sociaux des utilisateurs d'un système basé sur le

filtrage de l'information, ce qui leur donne le pouvoir d'avoir un contrôle direct pour filtrer un message indésirable (messages et commentaires) sur leurs murs. La plupart des recherches dans les médias sociaux ont été adressées aux classifications des tweets Liu, (2012), Al-Ayyoub et al. (2015), Contratres et al., (2018), Zhao et al., (2012), Kermanidis, and Maragoudakis, (2013), Tang et al., (2014). Seules quelques études ont porté sur la classification des messages de Facebook. Akaichi, et al. (2013), a réalisé une classification de sentiment sur une collection de données qui est une collection des statuts des utilisateurs tunisiens sur Facebook. Cette collection conduit à la révolution tunisienne. En concluant une procédure expérimentale comparative entre les classfiers Naïf Bayes et SVM, tandis que ces algorithmes peuvent atteindre une grande précision. Faqeeh et al. (2014) ont travaillé sur la Classification des documents de court textes dans un cadre inter-langues : arabe et en anglais, et comparé les performances de certains des classificateurs de documents les plus populaires sur deux ensembles de données de courts commentaires sur Facebook. Quelques travaux dans les médias sociaux se concentrent sur le langage texto, le langage informel dans les réseaux sociaux. Kundi, et al., (2014) ont présenté un cadre sur des ensembles de données de microblogs pour la détection et la notation des mots écrits en langage texto pour l'analyse des sentiments. Ce cadre renforce le prétraitement de ces textes. Amiri, (2012) a met l'accent sur le langage texto des mines et des mots d'opinion urbaines et des phrases de base communautaire question répondeurs.

Dans cette partie, nous avons discuté un état de l’art sur de différentes techniques, le data mining, le text mining, l’apprentissage automatique, le traitement du langage naturel, et le traitement de texte. Par la suite nous avons décrit la tâche globale de la catégorisation de textes, tout en distinguant ses variantes et exposant les solutions connues de la littérature pertinente. Et enfin nous avons donné un aperçu sur les travaux réalisés dans le traitement et la classification des textes issus des réseaux sociaux.

Ceci représente une importante source d'information et un moyen afin de suivre les tendances, les nouvelles et les idées partout dans le monde. Bien qu'il soit important de déterminer les sujets des publications individuelles pour extraire les intérêts des utilisateurs des médias sociaux, cette tâche est non triviale, car les publications sont hautement contextualisées et informelles et ont une durée limitée.

1.5. La modélisation des intérêts des utilisateurs et catégorisation des utilisateurs

Avec la demande croissante d'applications personnalisées, l'extraction et la modélisation des intérêts des utilisateurs gagnent de plus en plus d'importance. Diverses sources d'informations ont été utilisées pour extraire et modéliser des paramètres qui représentent les intérêts des utilisateurs. Les médias sociaux sont devenus l'une des plateformes les plus populaires et les plus importantes pour le partage et la diffusion d'informations. Ces plates-formes sociales offrent aux utilisateurs un moyen non seulement de partager le contenu de leurs intérêts, mais également de mieux comprendre leurs activités quotidiennes. L'exploitation de ce contenu pour définir les intérêts des utilisateurs peut être utilisée pour la personnalisation d'une variété d'applications commerciales et non commerciales, telles que le marketing et la recommandation de produits.

Dans cette partie, nous étalons les modèles et les travaux liés aux intérêts des utilisateurs et au regroupement des utilisateurs pourtant les mêmes centres d'intérêts dans les médias sociaux dans la littérature.

1.5.1. Déduction et modélisation des intérêts des utilisateurs

Déduire les intérêts des utilisateurs (en anglais, inffering users' interests) de leurs activités dans l'espace des réseaux sociaux a été un sujet de recherche émergent ces dernières années. Zarrinkalam et al. (2016). L'inférence des intérêts des utilisateurs en analysant le contenu partagé dans les réseaux sociaux joue un rôle important à plusieurs applications telles que les recommandations personnalisées, en particulier dans les situations de démarrage à froid (cold start problem) Piao, and Breslin (2018). Malgré la popularité de la déduction des intérêts des utilisateurs aux seins des RS, il n'existe pas d'examen approfondi des stratégies de modélisation des utilisateurs permettant d'inférer les profils des intérêts des utilisateurs dans les RS Piao, and Breslin (2018). Dans le contexte de la recherche sur la déduction des intérêts des utilisateurs aux siens des RS, la plupart des études se sont concentrées sur l'exploitation *d'informations implicites* telles que le contenu partagé par les utilisateurs afin d'inférer des profils d'intérêts d'utilisateurs.

Les chercheurs Jiang and Sha (2015) ont distingué trois types d'approches d'inférence des intérêts des utilisateurs dans les RS : les approches exploitant le contenu textuel partagé dans les réseaux sociaux, les approches exploitant des connaissances sémantiques externes (par exemple, dbpedia ou freebase) et les approches qui ont étudié les intérêts dans les utilisateurs des réseaux sociaux en utilisant plusieurs sources.

Suivant le premier type, beaucoup de travail a été fait pour détecter les intérêts des utilisateurs à partir de leurs publications sociales, ce qui est très important pour la construction

d'un modèle d'utilisateur précis, dans lequel en tentant de déduire les sujets d'intérêt à partir du contenu partagé dans les réseaux sociaux Hong et al. (2013); Xu et al. (2011); Kang et Lee (2017); Liu et al. (2018); Lin et al. (2017). Le deuxième type d'approche s'intéresse à l'exploitation de la structure des liens des sources de connaissance. Abel et al. (2011), Tao et al. (2011); Jiang et al. (2015); Piao et Breslin (2016); Zarrinkalam et al. (2015); Faralli et al. (2017); Shah et al. (2018). Autres travaux comme, Khemmarat et al. (2014); Orlandi et al. (2012) dans le troisième type ont étudié l'intérêt des utilisateurs au sein des RS en fusionnant de différentes sources.

Dans Abel et al. (2011), les auteurs ont évalué les profils d'utilisateur à court et à long terme et ont enrichi la sémantique des activités de micro-blogging pour la création de profils d'utilisateur sémantiquement reliant des messages Twitter à des articles de presse connexes. Kang et Lee (2017) ont associé le contenu des textes des médias sociaux aux catégories pertinentes des médias. Le travail de Xu et al. (2018), a révélé la tendance en surveillant le comportement de retweet via l'intérêt des utilisateurs et des catégories de tweets en modélisant la distribution par sujet des tweets. Dans Zarrinkalam et al. (2016), les auteurs ont proposé un schéma de prédiction de lien basé sur un graphique pour modéliser le profil d'intérêt d'un utilisateur sur un ensemble de sujets / événements présents sur Twitter dans un intervalle de temps spécifié. D'autres travaux ont utilisé Wikipedia pour modéliser un sujet d'intérêt émergent sur Twitter en conjonction de plusieurs concepts sémantiques de Wikipédia dans Zarrinkalam et al. (2015), et pour générer un graphe d'intérêt hiérarchique pour les utilisateurs de Twitter en tirant parti du graphe de catégories de Wikipedia dans Faralli et al. (2017). Quand d'autres traveaux Piao et Breslin (2016), Manrique et al. (2017) ont exploité les informations DBpedia pour modéliser les intérêts des utilisateurs. Dans Tao et al. (2011), les auteurs ont présenté TUMS, un service de modélisation utilisateur basé sur Twitter qui génère des profils utilisateur sémantiques basés sur des fonctionnalités permettant d'enrichir le sens des tweets. Les TUMS produisent des informations par entité, par sujet ou par profils basés sur les hashtags. Dans le travail de Xu et al. (2011) un modèle auteur-sujet a été proposé pour découvrir les sujets des utilisateurs d'intérêt sur Twitter.

Conscients que les intérêts des utilisateurs changent avec le temps, des travaux antérieurs ont examiné la dynamique temporelle présentant un intérêt pour la modélisation des utilisateurs dans les réseaux sociaux en ligne Abel et al. (2011) ; Budak et al. (2014); Jiang et al. (2015); Piao et Breslin (2016); Zarrinkalam et al. (2015); Shah et al. (2018); Zhang et al. (2017).

Dans cette thèse, nous traitons du premier type d'approche, en déduisent les sujets d'intérêts en analysant le contenu partagé dans les RS, tenant compte de l'importante caractéristique de la

dynamique temporelle dans l'inférence et la modélisation des intérêts des utilisateurs.

1.5.2. Catégorisation des utilisateurs

Dans les réseaux sociaux, les personnes sont interconnectées en formant des groupes d'utilisateurs (communautés). Par conséquent, le regroupement d'utilisateurs partageant des intérêts similaires est une tâche essentielle pour plusieurs applications, telles que les systèmes de suggestion d'amis, le filtrage collaboratif, la segmentation de la clientèle et l'analyse d'influence sociale.

Ce problème a été traité par plusieurs travaux, y compris, Moosavi et al. (2016); AlSuwaidan et Ykhlef (2017); Yamak et al. (2018); Abdelsadek et al. (2018); Madan et al. (2018); Zhao et al. (2016), ce problème est concédé comme un problème de distribution de graphes, dans lequel les utilisateurs sont représentés par des nœuds et les relations entre eux par des arêtes. En général, les relations représentent des liens d'amitié explicites, où les utilisateurs n'ont que peu de liens. De plus, au sein d'un même réseau social, les utilisateurs sont très éloignés les uns des autres par rapport à la vaste distribution des réseaux sociaux. Or, trouver et regrouper les utilisateurs en fonction de leurs intérêts communs est une tâche difficile et fastidieuse, ce qui montre la limitation de l'approche basée sur les liens. En outre, d'autres travaux, Jaffali et al. (2014a); Li et al. (2015); Ho et Do (2015); Jaffali et al. (2014b); Jarwar et al. (2017); Huang et al. (2018) ont proposé une autre méthode basée sur l'analyse des posts textuels pour regrouper les utilisateurs partageant les mêmes intérêts.

1.5.3. Conclusion

Après avoir énuméré les différents points liés à l'état de l'art, nous proposons un modèle générique qui couvre tous ces points de la littérature, et qui permet de constituer la structure de base pour la création de plusieurs types d'applications. Tous les points couverts par le modèle proposé sont présentés dans le chapitre suivant, avec une description détaillée.

Chapitre II

Méthodologie

2. Chapitre 2 : Méthodologie

2.1. Introduction

Ce chapitre représente notre méthodologie, nous avons réuni les deux différentes disciplines discutées dans le chapitre précédent (le text mining et les modèles sémantiques). En réutilisant l'ontologie SIOC pour modéliser les utilisateurs et leur contexte (profilage). Les techniques du text mining sont utilisées pour analyser et modéliser le contenu généré par les utilisateurs afin de capturer le sens de ce contenu. En autre terme, enrichir le profil utilisateur par le sens du contenu partagé. Cet enrichissement, nous permettrons par la suite de déduire et modéliser les intérêts dynamiques temporelle des utilisateurs et de regrouper ces utilisateurs selon leurs intérêts. En introduisant un modèle générique appelé le modèle de catégorisation de contenu et d'utilisateurs (en anglais, *Users and Content Categorization (UCC)*). Le reste de ce chapitre décrit le modèle proposé. En décrivant le processus d'utilisation et en détaillant tous les composants de ce modèle. Aussi, afin d'améliorer catégorisation de contenu, une nouvelle approche basée sur les attributs démographiques est proposée dans ce chapitre.

2.2. L'architecture du modèle UCC

Dans cette section, nous décrivons l'architecture de notre modèle. Cette architecture a pour le but d'aider les programmeurs d'applications à construire des modèles efficaces dans le contexte de leur domaine. Les entrées et les sorties (Inputs\Outputs) et les composants (les modules) de ce modèle sont montrées en Figure 2.1.

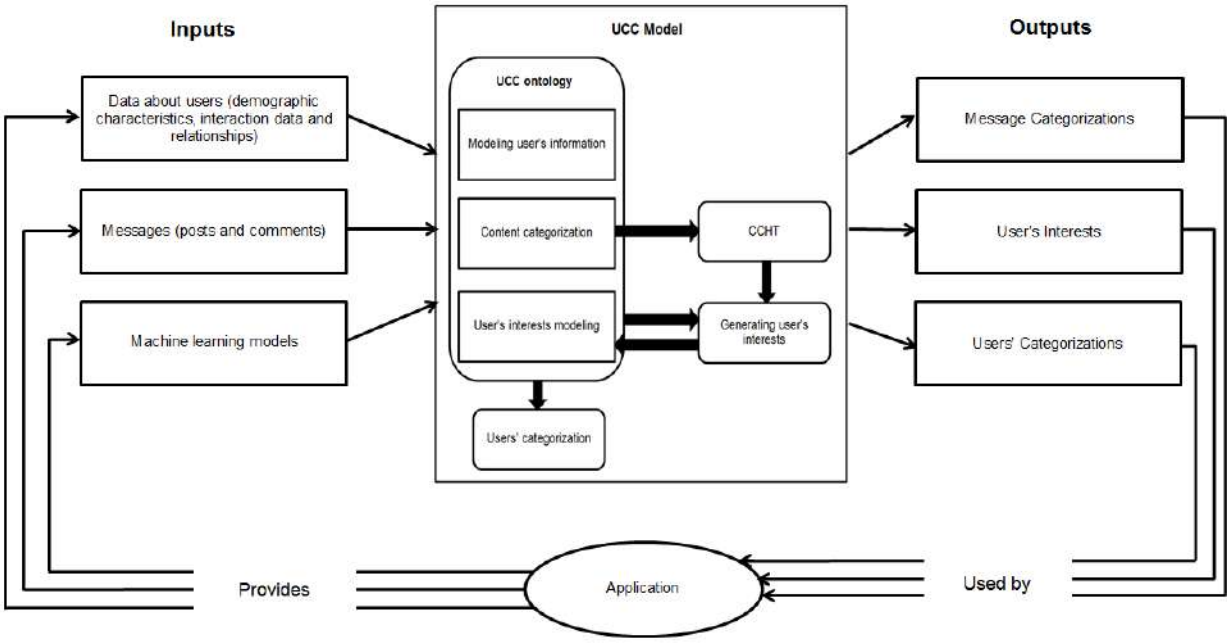

Figure 2.1: Le processus d'utilisation du modèle UCC.

Les entrées et les sorties de ce modèle sont :

- Les entrées (Inputs)
  - ✓ Les données des utilisateurs : elles représentent les données personnelles partagées par l'utilisateur (les attributs démographiques, telles que le nom, l'âge, le sexe, etc.), les données d'interactions, telles que (le contenu partagé : les publications et les commentaires (le contenu textuel), etc.) et la relation avec les autres utilisateurs (amitiés, relations familiales, etc.)
  - ✓ Les messages (publications et commentaires): ils représentent les données textuelles issues du contenu partagé par les utilisateurs.
  - ✓ Les modèles d'apprentissage automatique : ce sont les modèles construits lors du processus d'apprentissage automatique supervisé (la classification) avec tous ses composants (prétraitement du texte, les catégories, l'algorithme, etc.). Les catégories choisies et construites par l'application qui utilisera ce modèle. Le texte peut être catégorisé en utilisant de nombreuses classifications.
- Les sorties
  - ✓ Les intérêts des utilisateurs : ce sont le centre d'intérêt pour chaque utilisateur après la classification de ses messages partagés.
  - ✓ Catégorisation des messages : elle représente la classification du message (publication / commentaire) dans une ou plusieurs catégories en utilisant les modèles d'apprentissage d'automatique.
  - ✓ Catégorisation des utilisateurs : après avoir eu une idée complète sur les informations personnelles des utilisateurs et de leurs intérêts, le modèle attribut des catégorisations des utilisateurs en fonction de leurs intérêts.
- L'application : elle représente l'utilisateur final du modèle UCC, par exemple, un système de recommandation ou une détection de communautés en ligne. Notre modèle permet d'inclure l'utilisateur final dans le processus de développement d'un système.

Les sections suivantes décrivent en détail les déférents modules constituant le modèle UCC et les relations entre eux.

2.3. L'ontologie UCC

Pour atteindre notre objectif, nous avons développé l'ontologie UCC, qui fournit une description sémantique pour les réseaux sociaux. En enrichissant les vocabulaires existants (FOAF, SIOC), en ajoutant une couche, qui exprime le sens du texte partagé dans les RS, par intégration d'autres concepts concernant la classification du contenu textuel et les intérêts d'utilisateurs.

Le vocabulaire SIOC a pour objectif d'exprimer des informations sur la structure et le contenu des sites communautaires en ligne et de permettre l'interopérabilité sur le Web social à l'aide des technologies du Web sémantique.

2.3.1. Modélisation des profils utilisateurs

Aujourd'hui, l'ontologie SIOC est l'ontologie la plus complète et générique, développée non seulement pour capturer les caractéristiques des communautés en ligne, mais aussi pour modéliser le contenu partagé par ces communautés. Pour cette raison, notre modèle réutilise l'ontologie SIOC pour modéliser les profils utilisateurs et ces interactions dans les communautés en ligne.

✓ **L'ontologie SIOC**

Les principales classes de l'ontologie SIOC sont schématisées dans la Figure 2.2.

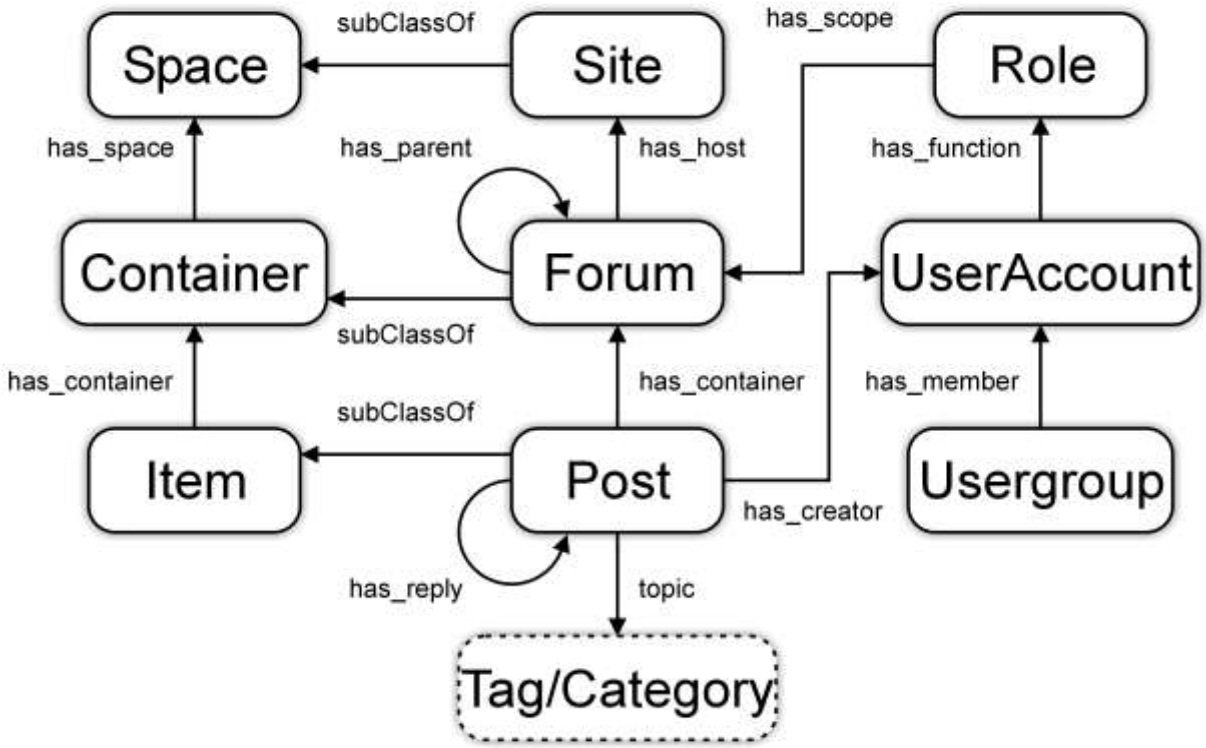

Figure 2.2: L'ontologie fondamentale du SIOC Breslin et al., (2005).

- *sioc:Community* : cette classe représente la communauté des utilisateurs en ligne.
- *sioc:Container* : cette classe définit une classe générique de la classe sioc:Space (sub-element), à savoir un forum, un thread, etc.
- *sioc:Forum* : cette classe représente la définition d'un forum spécifique, dérivée de la classe générique sioc:Container.
- *sioc:Item* : elle représente une classe générique définissant un document au sens général, c'est-à-dire un message de forum, un article de wiki, etc.
- *sioc:Post* : cette classe définit un post spécifique, pour représenter un post ou un commentaire dans un forum, elle est dérivée de la classe générique Sioc:Item.
- *sioc:Role* : cette classe représente le rôle d'un utilisateur (administrateur, invité, etc.).
- *sioc:Site* : cette classe spécifique définit un site dérivée de la classe générique sioc:Space, qui définit une classe générique d'un espace de travail (classe racine). Elle peut être un site web, un carnet d'adresses, un calendrier, etc.
- *sioc:Thread* : cette classe est une classe spécifique dérivée de la classe sioc:Container qui définit un thread.
- *sioc:UserAccount* : elle représente le compte d'utilisateur, c'est-à-dire une personne ayant déjà une définition ailleurs (FOAF, etc.).
- *sioc:Usergroup* : elle représente un groupe d'utilisateurs.

Afin de modéliser l'utilisateur, ses caractéristiques et ses relations avec les autres utilisateurs au sein de la communauté en ligne, la classe sioc:UserAccount de SIOC a les propriétés suivantes:

- ✓ *sioc:id* se réfère à l'ID de l'objet.
- ✓ *sioc:creator_of*, qui relie l'utilisateur au contenu qu'il génère.
- ✓ s*ioc:has_function* cette propriété renvoie le rôle de la personne dans l'espace de travail.
- ✓ *sioc :account_of* se réfère à la description de la personne en question. Ici, les descriptions FOAF pour la personne.

Cette classe réutilise les propriétés d'autres vocabulaires, comme :

- ✓ *foaf :nom* capture le nom de l'utilisateur,
- ✓ *foaf :age* capture l'âge de l'utilisateur,
- ✓ *dc:created*, capture l'heure et la date de création du compte d'utilisateur,
- ✓ *foaf:knows*, relie l'utilisateur à un autre utilisateur ou un utilisateur avec son réseau social.

Pour modéliser le processus de la création de contenu et l'interaction de l'utilisateur avec les autres membres de la communauté, SIOC utilise les classes *sioc:Item* et la classe spécifique *sioc:Post*, qui a la propriété:

- ✓ *sioc:has_creator* relie le message partagé avec un compte de l'utilisateur qui a posté le message.
- ✓ *sioc:hasParent* relie le post avec un thread particulier.
- ✓ *sioc:reply_of* et *sioc:has_reply*, indiquent que ce message est une réponse à un autre message.
- ✓ *sioc:attachment* est utilisé lorsqu'un document est attaché au message.
- ✓ *sioc:content* représente le texte du contenu, le corps du message, peut être soit une ressource externe ou le contenu inclus dans le RDF.
- ✓ *sioc:created* indiquer la date / l'heure, de publication du post.

La classe *Sioc:Thread* est également liée à la classe *Sioc:Container* (forum, blog, etc.) par la propriété *Sioc:has_parent.*

Par exemple, un Forum représente un espace dans lequel la discussion se produise (pas nécessairement un tableau d'affichage, en dépit de son nom, mais par exemple comme un blog), et contient différents messages, rédigés par des utilisateurs. L'exemple suivant décrit ensuite qu'Alice a créé un post dans un forum particulier (une zone de discussion) et que Bob a répondu comme suit (en utilisant la notation N3 - préfixes omises). Afin de représenter des conteneurs plus abstraits (comme un espace d'information personnelle), des classes plus générales (Container et Space) peuvent être utilisés.

```
:post a sioc:Post ;
sioc:has_creator :alice ;
sioc:has_container :forum ;
sioc:has_reply :reply .
:forum a sioc:Forum .
:reply a sioc:Post ;
sioc:has_creator :bob .
:alice a sioc:User .
:bob a sioc:User .
```

Après la modélisation de l'utilisateur, ces données personnelles et sociales (profil d'utilisateur), nous allons montrer comment enrichir ce profil, en donnant un sens au contenu.

2.3.2. Enrichissement des profils utilisateurs

En réutilisant les classes et les propriétés SIOC, notre modèle capture les informations démographiques de l'utilisateur ainsi que ses partages (lorsque l'utilisateur poste un message, quand il répond, etc.) au sein des communautés en ligne.

Certaines propriétés concernant la création de contenu d'interaction représentent une connaissance essentielle pour certaines applications qui ne sont pas prises en compte par l'ontologie SIOC, telles que :

- Connaître le nombre de partages ou le nombre de likes sur un message est un élément élémentaire pour mesurer l'interaction des utilisateurs avec un sujet spécifique.
- Connaître la profession et la localisation des personnes permettent de bien cerner les caractéristiques qui intéressent à un sujet donné, par exemple, 62% des personnes qui ont des tendances terroristes (partagent, aiment ou commentent positivement les messages encourageant l'acte terroriste) se trouvent dans la région X, ou, ces personnes n'ont pas un niveau d'éducation élevé.
- Pour pouvoir manipuler le contenu textuel des réseaux sociaux, il est nécessaire de déterminer la langue dans laquelle il est écrit afin de choisir les outils de traitement du langage naturel.

Par conséquent, nous avons ajouté quelques propriétés à la classe sioc:Post. Ces propriétés sont les suivantes :

- ✓ *ucc:num_of_share* représente le nombre de fois que cette publication est partagée.
- ✓ *ucc: num_of_likes* représente le nombre de likes associés à ces publications.
- ✓ *ucc: has_language* fait référence à la langue utilisée pour écrire le contenu du post.
- ✓ *ucc: has_profession* fait référence à la profession de l'utilisateur.
- ✓ *ucc: has_location* fait référence à l'emplacement de l'utilisateur.

Après avoir modélisé les données sur les utilisateurs, les relations et les interactions entre eux, nous présentons la deuxième partie de notre ontologie, qui décrit la catégorisation des publications et la modélisation des intérêts de l'utilisateur. La catégorisation des messages contient les mêmes éléments que l'apprentissage automatique, une classe pour la classification, une classe pour les catégories de cette classification, une classe pour le prétraitement, une classe pour l'algorithme utilisé et une classe pour le modèle proprement dit. Les classes de la deuxième partie sont décrites à la Figure 2.3.

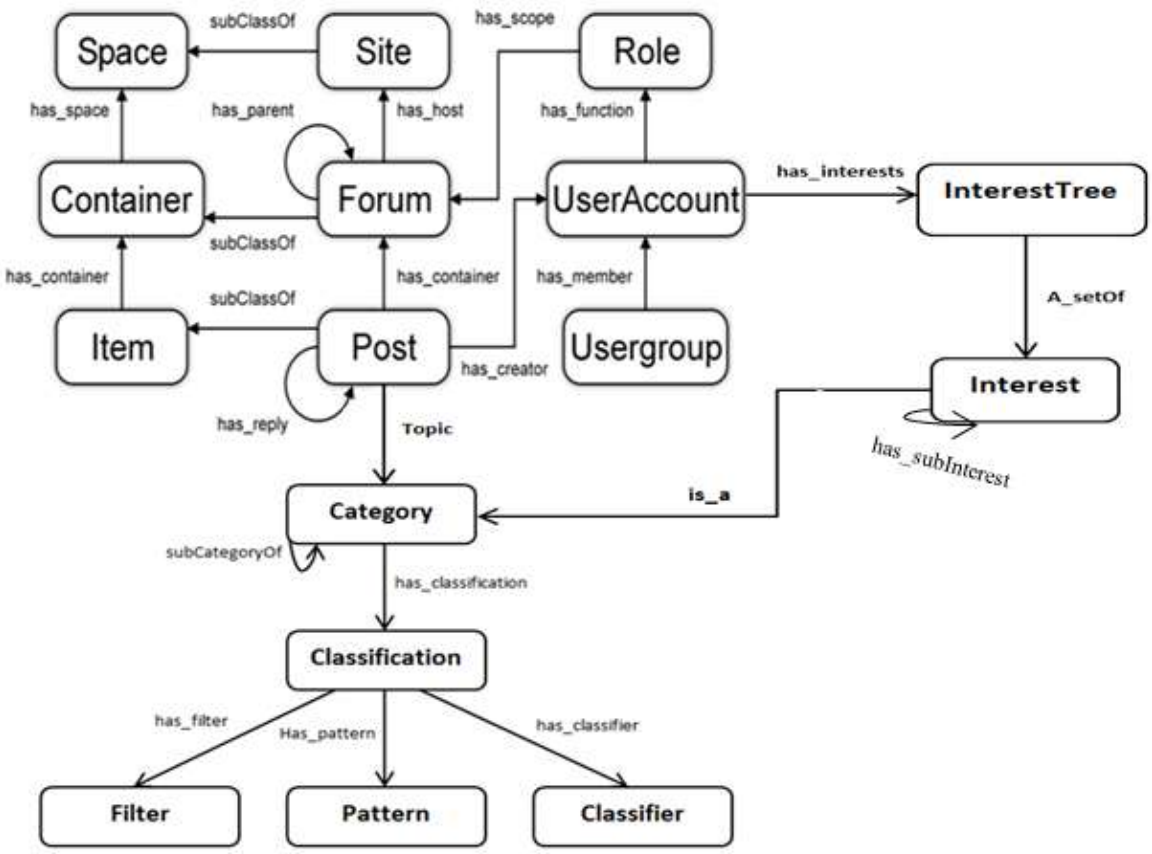

Figure 2.3: L'ontologie UCC

- ✓ *ucc:Catégory* représente la classe qui attribue la catégorie à la publication, dans laquelle une publication peut être étiquetée avec plusieurs catégories ou par catégories hiérarchiques. Cette classe est liée à *sioc:Post* par la propriété *sioc:topic.*
- ✓ *ucc:Classification* représente la classification attribuée par l'application, par exemple la classification des sentiments. La propriété *has_classification* lie les catégories à leur classification.
- ✓ *ucc:Filter* représente l'étape de prétraitement du texte.
- ✓ *ucc:Pattern* définit le modèle construit pour cette classification.
- ✓ *ucc:classifier* définit les algorithmes utilisés pour cette classification, et qui sont liés à la classe *ucc:Classification* par les propriétés *has_filter*, *has_pattern*, and *has_classifier*.
- ✓ *ucc:InterestTree* représente l'arborescence des catégories d'intérêts des utilisateurs. Cette classe est liée à la classe *sioc:UserAccount* par la propriété *ucc:has_intérets.*
- ✓ *ucc:interest* est une *ucc:Category* dont l'utilisateur est intéressé. Cette classe est liée à la classe *ucc:InterestTree* par la propriété *ucc:a_setOf.* Il a la propriété *ucc:has_score*, qui représente le score associé à chaque intérêt pour chaque utilisateur, et a une période représentée par *has_startdate* et *has_enddate.* Un intérêt peut avoir un ou plusieurs sous-

intérêts en utilisant une propriété *has_subInterest.*

2.3.3. L'implémentation de l'ontologie

L'ontologie UCC est développée en utilisant le logiciel Protégé qui est un éditeur open source d'ontologie distribué par l'Université de Stanford Medical Informatics. Protégé est un éditeur très extensible, capable de gérer une grande variété de formats (RDF, OWL, etc.).

Notre ontologie est représentée dans OWL (Web Ontology Language), un langage formel basé sur XML qui fournit un moyen formel de décrire les concepts de domaine. Une prise d'écran de l'ontologie créée est illustrée à la Figure 2.4.

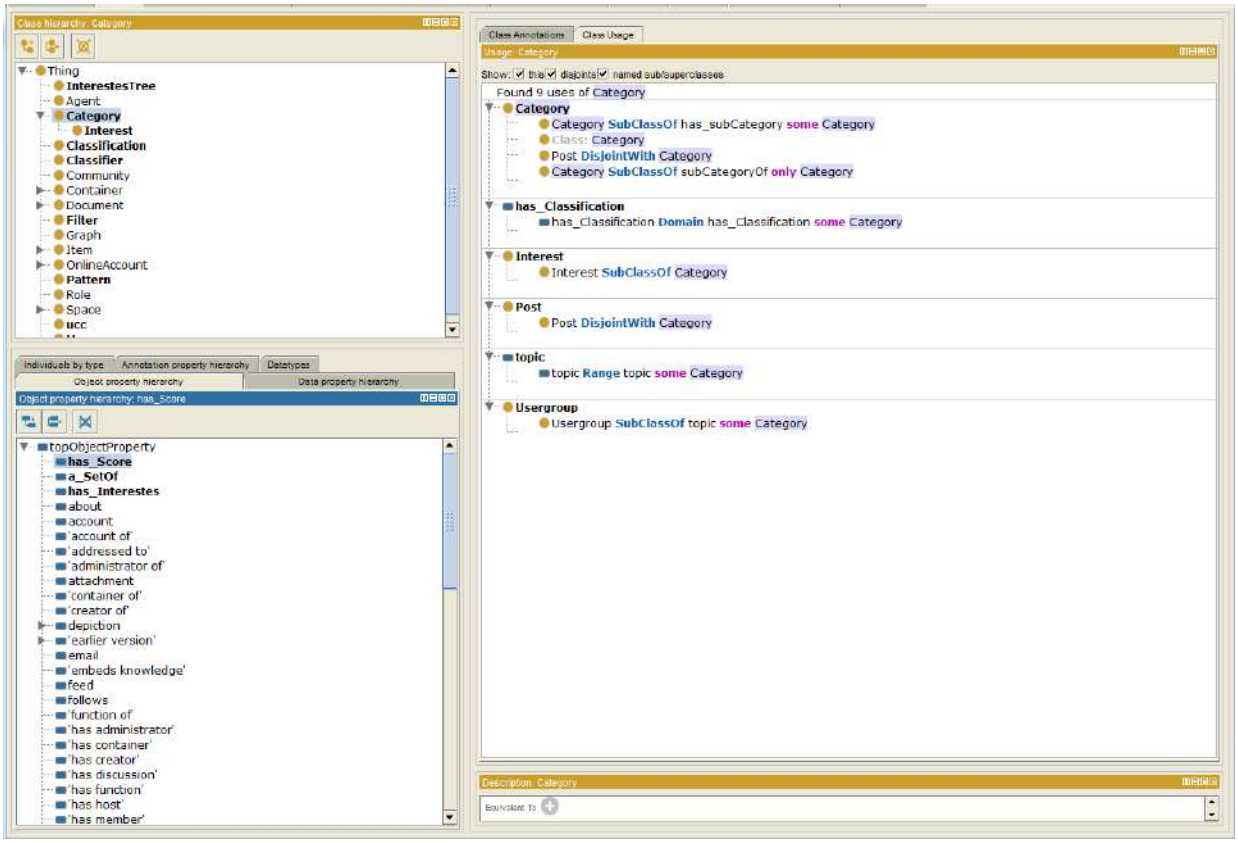

Figure 2.4 : Une prise d'écran de l'ontologie UCC.

Après avoir construit l'ontologie UCC, l'application peut choisir et construire l'arbre hiérarchique de catégorisation du contenu (en anglais, *Content's Categorization Hierarchical Tree (CCHT)*), en fonction des modèles d'apprentissage automatique. Les détails sont présentés dans la section suivante.

2.4. Arbre hiérarchique de catégorisation du contenu

Inspiré par la taxonomie de l'arbre hiérarchique des sujets (Topic Hierarchy Tree (THT)), développée par Jiang and Sha (2015). Nous construisons CCHT. THT est basée sur l'utilisation des sources de connaissances externes (par exemple, dbpedia, Freebase, yago) comme base de

connaissances pour sa construction. CCHT est construite en utilisant l'apprentissage automatique (la classification supervisée), elle peut supporter un grand nombre de catégories sélectionnées par l'application. Cette sélection peut inclure tous les aspects de ce que les gens parlent. CCHT définit des classifications et des catégories pour plusieurs niveaux. Le premier niveau est utilisé comme racine (afin de conserver l'aspect arbre). Le niveau 2 représente les classifications, y compris les catégories imbriquées sous la racine. Le niveau i représente les classifications et les catégories imbriquées selon les classifications et les catégories du niveau i-1 (la Figure 2.5. représente un exemple d'un CCHT).

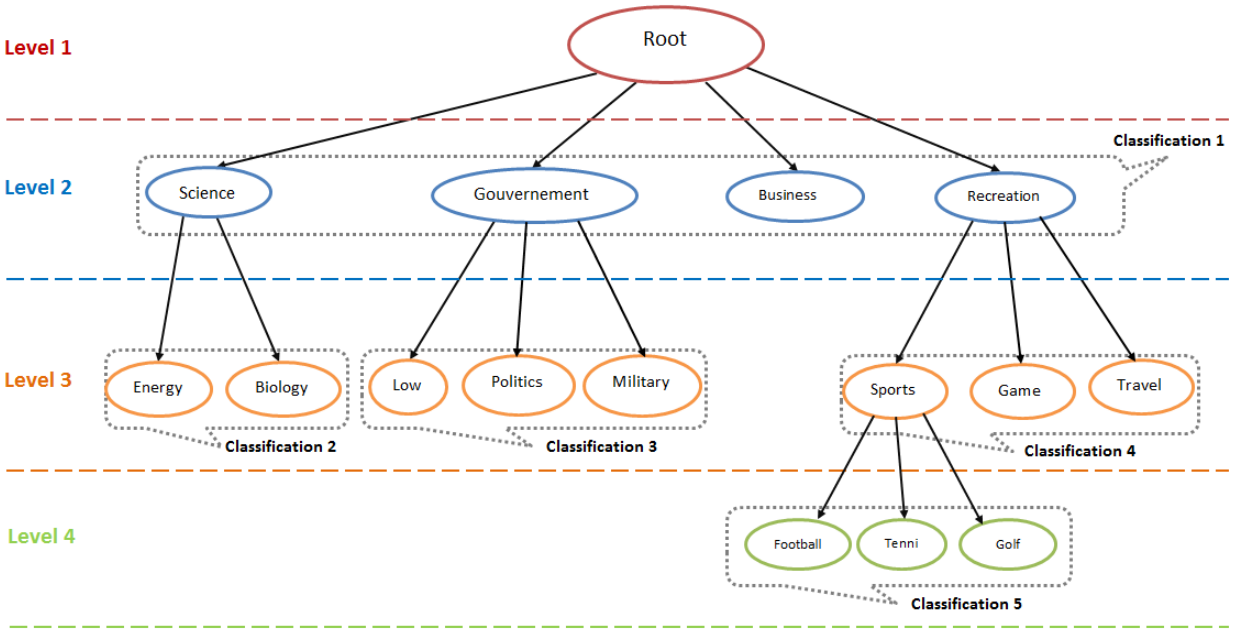

Figure 2.5 : un exemple d'un arbre hiérarchique de catégorisation du contenu.

Nous exploitons CCHT pour classer le contenu textuel partagé par les utilisateurs dans plusieurs niveaux jusqu'au sixième (6émé) niveau au maximum, de niveau plus général au plus précis. Les niveaux plus profonds permettent des classifications plus précises pour les messages, ce qui permet de mieux comprendre et clarifier les intérêts de l'utilisateur.

2.5. La déduction des intérêts des utilisateurs

Plusieurs travaux antérieurs ont focalisé sur le changent et évoluent des intérêts des utilisateurs avec le temps. Le modèle UUC traite ce problème en proposant de générer les intérêts d'un utilisateur selon la catégorisation des messages partagés par l'utilisateur. Par le biais d'un arbre d'intérêts marqué, où chaque intérêt est une catégorie qui est représentée par la classification d'apprentissage automatique, le score indique l'importance de l'intérêt pour un utilisateur donné. Nous considérons les intérêts de l'utilisateur dans une période spécifique *d*, comme un ensemble

d'intérêts *I* où chaque intérêt est une catégorie *c* avec un score *s* associé à cette catégorie dans cette période de temps *d*.

$$I_u(d) = \{(c,s(c,d))|c \in C\} \qquad (1)$$

Où, *C* est l'ensemble de catégories c_j, et *P* est l'ensemble de messages p_i partagé par l'utilisateur *u* dans la période de temps *d*.

$$z(p_i, c_j) = \begin{cases} 1, si\ p_i\ est\ classé\ dans\ c_j \\ 0, sinon \end{cases}$$

Où, $p_i \in P$ *et* $c_j \in C$

$$S_{u,c_j}(d) = \sum_i z(p_i, c_j) \qquad (2)$$

Où *S* est un score calculé pour la catégorie c_j en fonction des posts partagés par l'utilisateur *u* dans la période du temps *d*.

En fin, la construction de l'arbre des intérêts des utilisateurs (en anglais, *User Interest Tree (UIT)*) : cet arbre est un CCHT scoré, ou, chaque intérêt d'un utilisateur donné *u* a un score calculé est ajouté en tant que nœud à l'arbre des utilisateurs dans l'ontologie UCC. La Figure 2.6. montre un exemple d'un arbre d'intérêt d'un utilisateur User-A en novembre 2015.

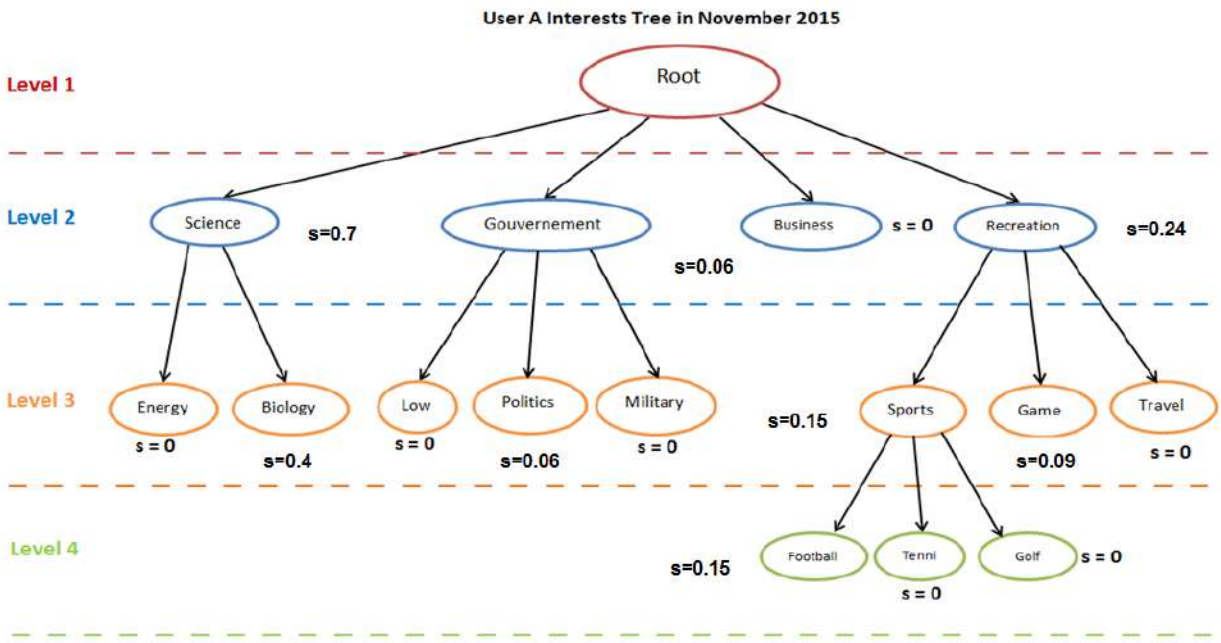

Figure 2.6: un exemple d'un arbre d'intérêt d'un utilisateur User-A en novembre 2015.

Les intérêts des utilisateurs peuvent être divisés en intérêt permanent et intérêt temporel. Intuitivement, l'intérêt permanent de l'utilisateur tient la stabilité au fil du temps ; Où, l'intérêt temporel est plus susceptible de suivre des sujets chauds ou des événements dans un moment précis.

Après avoir généré et modélisé les intérêts des utilisateurs, ces derniers peuvent être regroupés par catégories en fonction de leurs intérêts. La section suivante décrit les approches proposées pour la catégorisation des utilisateurs.

2.6. La catégorisation des utilisateurs

L'exploitation des intérêts et les activités des utilisateurs, modélisés dans le modèle UCC, peut améliorer la qualité de regroupement d'utilisateurs ayant des intérêts similaires au sein de la même catégorie.

L'objectif principal de cette section est de regrouper les utilisateurs en fonction de leurs intérêts extraits de la catégorisation des posts partagés par eux dans les réseaux sociaux pendant une période de temps d.

Nous avons proposé deux approches pour catégoriser les utilisateurs en fonction de leurs intérêts:

a- Le regroupement des utilisateurs en fonction de la ressemblance d'intérêts des utilisateurs-cœur

b- Le regroupement des utilisateurs en fonction de la catégorie d'intérêt

2.6.1. Le regroupement d'utilisateurs en fonction de la similarité d'intérêts avec les utilisateurs-cœur

Cette approche vise à regrouper les utilisateurs qui ont une forte similarité avec un utilisateur-cœur u_c dans une période de temps *d*. Les utilisateurs-cœur sont désignés par l'application (l'utilisateur final du modèle UCC). Cette approche repose sur les étapes suivantes :

1. Les entrées : {*setOfUsers* (L'ensemble de tous les utilisateurs), *setOfUsers-core* (lensemble des utilisateurs coeur), *d* (la pireode de temps), *minSim* (la valeur minimal de la similarité)}

2. Construire l'IUT pour tous les utilisateurs

3. Calculer la similarité entre chaque u_c et tous autres utilisateurs *u* et la comparer avec *minSim*

4. Si la similitude est supérieure ou égale à *minSim* alors ajouter *u* a la categorie C_{uc}

5. Sorties {C_{u_c} (afficher toutes les catégories)}

Toutes ces étapes sont fournies dans un pseudo-code de l'algorithme (voir, Algorithme 1)

Algorithm 1 : Grouping users based on users-core interests similarity approach
Input {setOfUsers, setOfUsers-core; d; minSim}
1. Foreach *Element u in setOfUsers* ***Do***
2. *Build UIT(u,d)*
3. End For
4. Foreach *Element* u_c *in setOfUsers-core* ***Do***
5. Foreach *Element u in setOfUsers* ***Do***
6. If *(Similarity (u,*u_c*) ≥ minSim)* ***Do***
7. *Said u Add to* C_{u_c}
8. End If
9. End For
10. ***End For***

Pour mieux comprendre ce procédé un exemple est fourni, dons le résultat est dans la Figure 2.7.

Exemple : nous avons un ensemble de 10 utilisateurs *setOfUsers* = {u_1, ..., u_{10}}, où les utilisateurs-coeur sont *setOfUsers-core* = {u_4, u_7} dans une période de temps *d* du *(2015-11-11 jusqu'au 2015-12-1)*, nous voulons extraire les catégories d'utilisateurs *C* = {C_{u_4}, C_{u_7}} qui regroupent les utilisateurs qui ont des intérêts similaires avec les utilisateurs-coeurs u4 et u7. Le résultat de cet exemple est illustré dans la Figure 2.7.

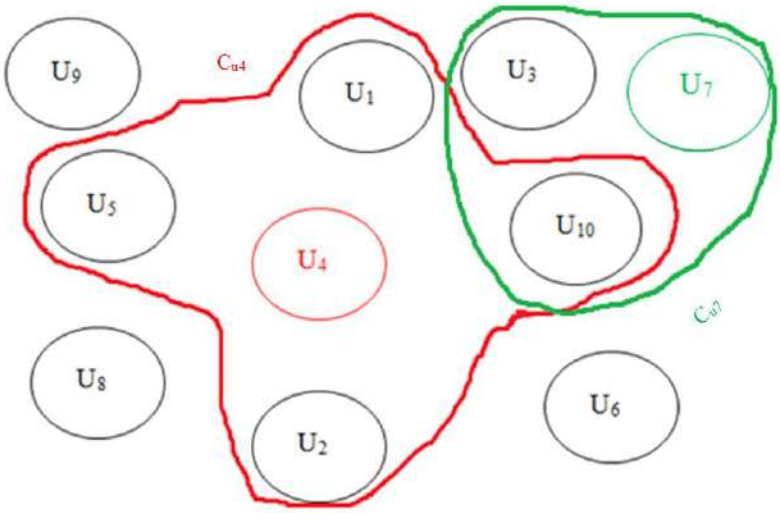

Figure 2.7: un exemple de regroupement des utilisateurs en fonction de la similarité d'intérêts avec les utilisateurs-cœur

2.6.2. Le regroupement des utilisateurs en fonction d'une catégorie d'intérêt

Cette approche vise à regrouper tous les utilisateurs intéressés à chaque catégorie *c* dans une période de temps *d*. interroger l'ontologie UCC permet de réaliser cet objectif. Il est trop important d'extraire les bonnes informations et de les utiliser correctement. C'est pourquoi nous avons choisi SQWRL (langage de règle Web amélioré par requête sémantique), qui est un langage de requête sémantiquement robuste pour que OWL communique l'ontologie UCC. SQWRL est construit sur la langue de règle SWRL OConnor et al., (2009). Il utilise la solide base sémantique de SWRL comme base formelle pour atteindre notre objectif.

Cette méthode vise à catégoriser les utilisateurs dans les réseaux sociaux, elle se compose de trois étapes principales : pour mieux comprendre cette méthode, nous fournissons un exemple où le résultat est montré dans la Figure 2.8.

Exemple : nous avons un ensemble de *10* utilisateurs $U = \{u_1, \ldots, u_{10}\}$, et un ensemble de 4 catégories $C = \{c_1, c_2, c_3, c_4\}$, dans une période de temps de *(2015-11-11 à 2015-12-1)*, nous voulons extraire les intérêts de chaque utilisateur, où le score s>= 0.10 et regrouper ces utilisateurs en fonction de la catégorie d'intérêts.

- ✓ Etape 1 : Extraire les utilisateurs isolés intéressés par chaque catégorie c en utilisant la requête sqwrl :

```
Person(?p) ^ name(?p; ?n) ^ UserAccount(?u) ^ account(?u; ?p)
^InterestT ree(?t)^Interest(?i)^hasınterests(?u; ?t)^a setOf(?t; ?i)
^has startdate(?t; ?d1) ^ has enddate(?t; ?d2) ^ has score(?i; ?s)
^swrlb : greaterThan(?s; 0.10) ^ swrlb : equal(?d1; 2015 - 11 - 11)
^swrlb : equal(?d2; 2015 - 12  - 11) ! query : select(?n; ?i; ?s)
```

- ✓ Étape 2 : Construire le fichier des utilisateurs, ce fichier représente les résultats obtenus en appliquant la requête SQWRL de l'étape précédente, voir le Tableau 2.1.

L'utilisateur	L'intérêt	Score
u1	C1	0.16
u1	C4	0.21
u2	C3	0.11
u3	C1	0.18
u4	C2	0.17
u4	C3	0.18
u4	C4	0.27
u5	C4	0.32
u6	C2	0.56
u7	C1	0.12
u8	C3	0.42
u9	C4	0.22
u10	C1	0.19
u10	C2	0.63

Tableau 2.1 : Le fichier des utilisateurs.

- ✓ Etape 3 : Construire les groupes d'utilisateurs, le résultat est affiché dans le Tableau 2.2.

La catégorie	**Le regroupement d'utilisateurs**
C1	G1={u1,u3,u7,u10}
C2	G2={u4,u6,u10}
C3	G3={u2,u4,u8}
C4	G4={u1,u4,u5,u9}

Tableau 2.2 : La construction des groupes d'utilisateurs

La Figure 2.8 représente le résultat obtenu en appliquant cette méthode sur cet exemple.

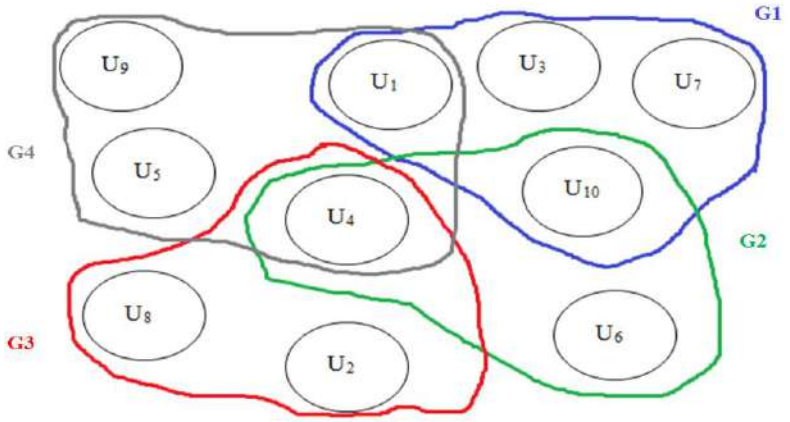

Figure 2.8 : un exemple de regroupement des utilisateurs en fonction d'une catégorie d'intérêt

Comme nous avons déjà montré, l'un des éléments de base du modèle UCC est la catégorisation du contenu en utilisant l'apprentissage automatique. Ou, les modèles d'apprentissage automatique doivent êtres puissant avec un minimum de taux d'erreurs. Afin de fournir un modèle puissant, nous proposons une nouvelle approche pour améliorer les modèles d'apprentissage automatique afin d'avoir une catégorisation de contenu plus performante. La section suivante décrit en détail l'approche proposée.

2.7. Exploitation des informations démographiques dans la classification du texte social

Les attributs démographiques des utilisateurs peuvent fournir certaines préférences pour certains sujets d'intérêt. Les personnes ayant différents attributs démographiques peuvent être intéressées par différents sujets, alors que les personnes ayant des attributs démographiques similaires peuvent avoir les mêmes intérêts. D'habitude, les jeunes s'intéressent à la technologie qu'aux personnes âgées, qui s'intéressent davantage à l'actualité politique beaucoup plus que les jeunes.

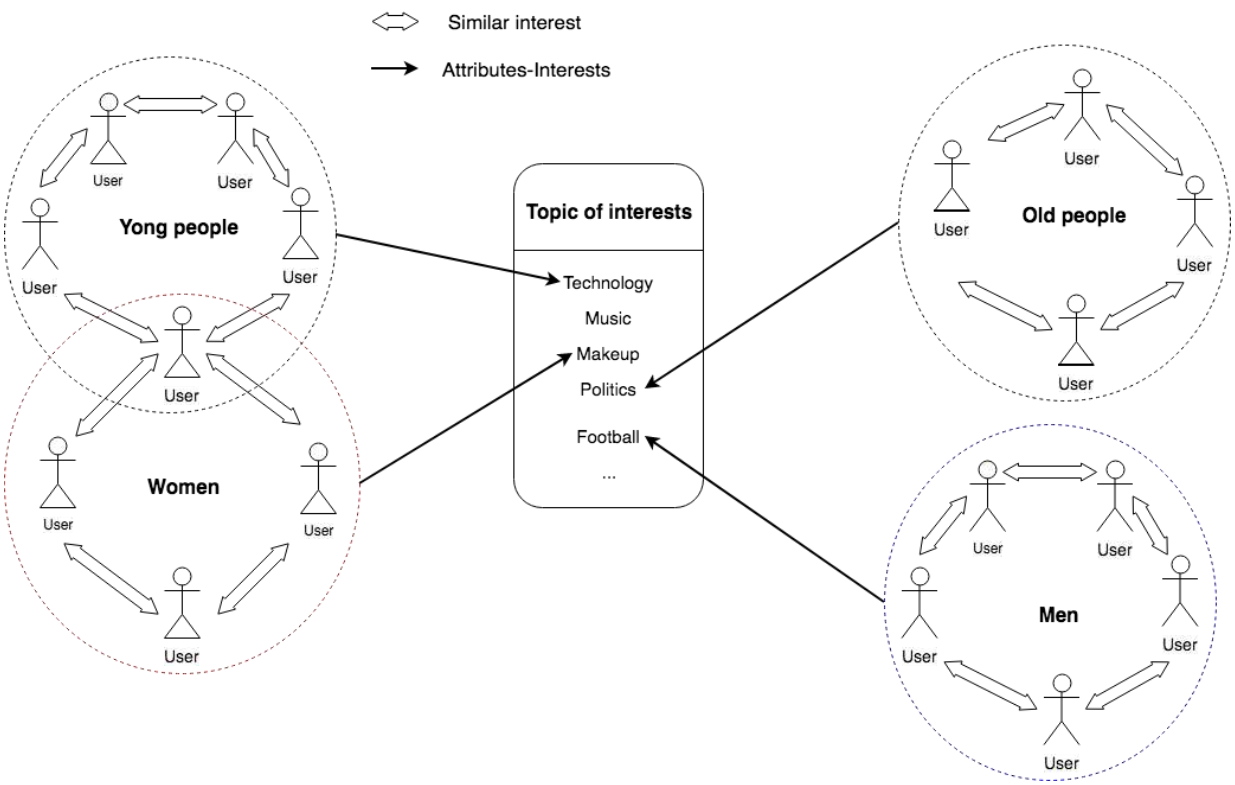

Figure 2.9: une illustration pour expliquer la relation entre attributs et intérêts des utilisateurs Sujet des intérêts et intérêts similaires.

- ✓ Sujet d'intérêts : les attributs démographiques des utilisateurs peuvent fournir certaines préférences pour certains sujets d'intérêt ; les personnes ayant des attributs différents peuvent ne pas s'intéresser aux mêmes sujets.
- ✓ Intérêts similaires : les personnes ayant des attributs similaires peuvent partager des publications ayant les mêmes intérêts. Habituellement, les personnes ayant un niveau d'éducation élevé sont plus intéressées par les articles scientifiques que par les autres.

Malgré le nombre considérable d'utilisateurs dans les réseaux sociaux, les informations démographiques sont disponibles avec une grande diversité de contenus générés par les utilisateurs. Toutes les études précédentes sur la classification du texte dans les RS exploitent uniquement les données textuelles extraites des interactions des utilisateurs, en ignorant et négligeant les attributs démographiques des utilisateurs, ce qui permettrait d'améliorer les performances de la classification. Par conséquent, il est possible de d'exploiter ces attributs pour créer des modèles d'apprentissage plus intelligents en obtenant de meilleures performances.

Dans ce travail, nous étudions comment le contenu généré par l'utilisateur (données textuelles) et les attributs personnels sont exploités pour catégoriser le contenu textuel par sujets.

Afin de comprendre le lien entre les attributs démographiques des utilisateurs et le texte généré par les utilisateurs, nous proposons une analyse quantitative, dont les résultats apparaissent au niveau du chapitre suivant.

Ici, nous faisons une approche basée sur le contenu et les attributs démographique (en anglais, A Demographic-Content-based approach) pour classifier le contenu textuel par sujets d'intérêt en tenant compte à la fois des caractéristiques textuelles et des caractéristiques démographiques dans la classification.

- **L'approche proposée**

  En général, dans la classification de texte, les classifiers d'apprentissage automatique sont des méthodes basées sur le contenu, où le vecteur des caractéristiques d'entrée $X = (x_1, x_2, \ldots, x_m)$ d'un classifier sont l'ensemble des mots construisant un document. Dans les méthodes basées sur le contenu, les classifiers n'utilisent que le contenu textuel pour affecter une classe c à un post p_i.

  Dans ce travail, notre objectif est d'apprendre aux classifiers d'utiliser le contenu textuel de la publication (les mots), et d'utiliser également les attributs démographiques de l'auteur de cette publication. La Figure 2.10. montre le processus de la classification en utilisant l'approche proposée qui montre comment intégrer les attributs démographiques au processus de classification.

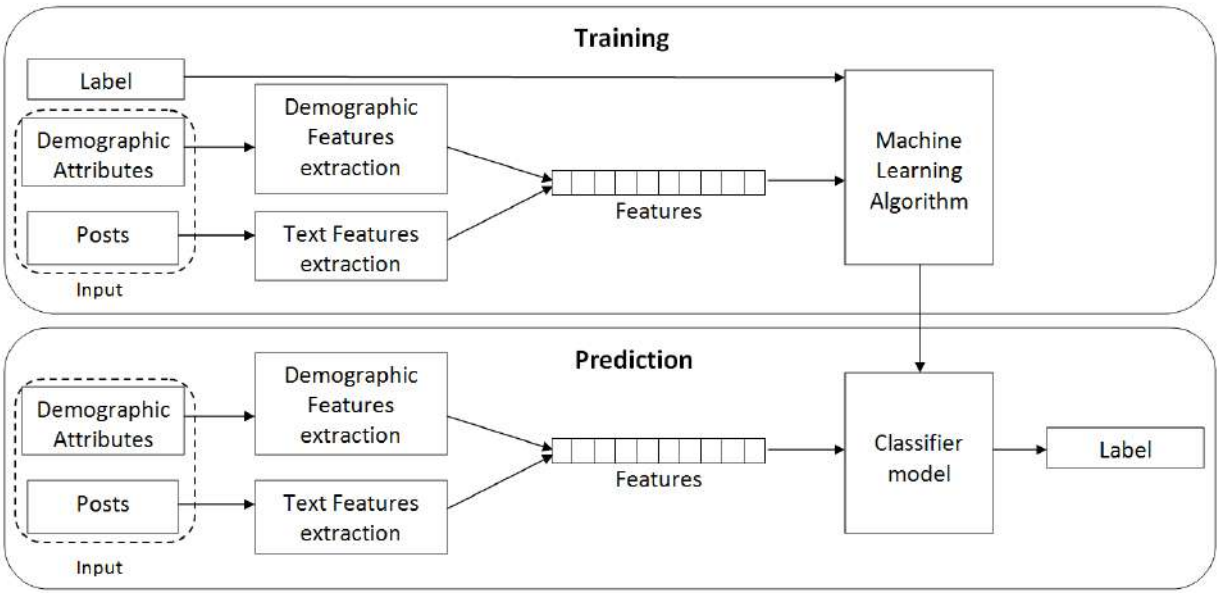

Figure 2.10:le processus de la classification de l'approche basée sur le contenu et les attributs démographique de l'utilisateur.

Comme le montre la Figure 2.10, les attributs démographiques sont intégrés comme entrées au classifier.

Maintenant, nous expliquons comment intégrer les attributs démographiques dans chaque classifier. Nous utilisons six classifiers très connus parmi les classifiers les plus utilisés dans la classification de texte, descendant de différentes approches de conception. Les classifiers considérés dans ce travail sont les suivants : (Articial Neural Network, K Nearest Neighbors,

Nave-Bayes, Decision Tree, Decision Rules and Support Vector Machine).

- **Artificiel Neural Networks (ANN)**

  ANN se compose d'éléments de calcul (neurones) fortement connectés les uns aux autres Bilski, (2011). Un neurone formel est une fonction algébrique paramétrée, avec des valeurs liées, de variables réelles appelées les entrées.

  En règle générale, le calcul de la valeur de cette fonction peut être décomposé en mappant la somme des poids *w* des entrées *d* dans la sortie *s* en fonction de la fonction d'activation *f* (telle que linéaire, tangente hyperbolique ou sigmoïde).

  En général, les résultats de la classification ANN basée sur le contenu sont les suivants :

$$s = f\left(\sum_{j=1}^{m} w_j * x_j\right) \qquad (3)$$

Maintenant, nous introduisons la sortie de la classification en utilisant ANN en se basant sur le contenu et les attribues démographique, pour chaque post p_i, le vecteur *V* est le vecteur correspondant au contenu et aux attributs démographiques.

$$s = f\left(\sum_{j=1}^{k} w_j * v_j\right) \qquad (4)$$

Dans la classification de texte, l'utilisation du ANN permet de manipuler de grands ensembles de caractéristiques et de corriger la classification en présence de bruit. Bilski, (2011).

- **K Nearest Neighbors (K-NN)**

  Le classifier k-NN trouve les K-voisins les plus proches parmi les documents d'apprentissage et utilise les classes des K-voisins les plus proches pour pondérer les candidats des classes. Le score de similarité de chaque document voisin le plus proche par rapport au document de test est utilisé comme poids des classes du document voisin. Formellement, la règle de décision dans la classification k-NN en basant sur le contenu peut s'écrire comme suit :

$$score\ (p_i, c) = \sum_{p_k \in KNN(p_i)} Sim\ \ (d_i, d_k)\ \delta\ (p_k, c) \qquad (5)$$

  En haut, $KNN\ (p_i)$ indique l'ensemble des k-voisins les plus proches du post p_i. $\delta\ (pk, c)$ est la classification du post p_k dont son contenu est d_k, par rapport à la classe *c*, c'est-à-dire:

$$\delta(p_k, c) = \begin{cases} 1, p_k \in c \\ 0, p_k \notin c \end{cases}$$

  La classe qui a la somme poids la plus élevée est affectée au post p_i.

  Maintenant, nous incluons l'attribut démographique dans la règle de décision Classification KNN, nous obtenons :

$$score\ (p_i, c) = \sum_{p_k \in KNN(p_i)} Sim\ \ [(d_i, A_u), (d_k, A_W)]\ \delta\ (p_k, c) \qquad (6)$$

- **Naïve Bayes (NB)**

  Le modèle probabiliste des classifiers NB est basé sur le théorème de Bayes et l'adjectif naïve provient de l'hypothèse selon laquelle les entités d'un ensemble de données sont mutuellement indépendantes. En pratique, l'hypothèse d'indépendance est souvent violée, mais les classifiers NB fonctionnent toujours bien sous cette hypothèse irréaliste.

  $$P(c/ p_i) = P(c/ d_i) \qquad (7)$$

  où $P\ (c/p_i)$ peut être calculé avec n'importe quel modèle génératif ou discriminatif.

  $$P(c/ p_i) = P(c/(d_i ,A)) \qquad (8)$$

- **Decision Tree (DT)**

  Habituellement, en utilisant l'approche basée sur le contenu pour classer un post p_i, un classifier DT place le contenu du post d_i à la racine et le propulse dans toute la structure de requête, t_n représente un certain test lié à un nœud *o*. *o [r]* est un nœud de progéniture ou une feuille de progéniture à laquelle une branche d'un nœud *o* mène. Cette branche correspond au résultat *r*. Le processus se poursuit ainsi jusqu'à atteindre une feuille représentant une catégorie particulière *c*.

  Dans l'approche proposée basée sur le contenu démographique pour la classification d'un poste, un classifier DT saisit à la fois le contenu d_i et les attributs démographiques A_u.

- **Decision Rules (DR)**

  Ce type de classification est basé sur des règles de déduction. Pour chaque catégorie *c*, l'algorithme crée un ensemble de règles correspondant au profil de cette catégorie. Les règles sont généralement construites comme suit :

  IF condition THEN conclusion

  Dans l'approche basée sur le contenu, la partie conditionnelle est basée sur les caractéristiques textuelles représentatives de la publication liée à une certaine catégorie, tandis que la partie conclusion représente la catégorie.

  Ici, nous introduisons une approche basée sur le contenu et les attributs démographiques, dans laquelle la partie conditionnelle est basée à la fois sur les caractéristiques textuelles et les attributs démographiques pour chaque catégorie c.

- **Support Vector Machine (SVM)**

  La plupart des problèmes de catégorisation du texte sont linéairement séparables Joachim, (1998). Un SVM Dumais, et al. (1988) sous sa forme linéaire, est une classe discriminante formellement définie par un hyper-plan qui sépare un ensemble d'exemples positifs d'un ensemble d'exemples négatifs avec une marge maximale. Pour chaque poste p_i, le vecteur de

contenu correspondant d_i est basé uniquement sur le contenu textuel du message. La formule pour la sortie d'un SVM linéaire est la suivante :

$$f(p_i) = \vec{w} \cdot \vec{d_i} + b \qquad (9)$$

où $\vec{w}$ est le vecteur normal de l'hyperplan et $\vec{d_i}$ est le vecteur d'entrée. La classe du post est attribué par le signe de :

$$h = sign(f(p_i)) = \begin{cases} +1, si\ f(p_i) > 0 \\ -1, sinon \end{cases}$$

Pour les modèles SVM multi classes, une classification à plusieurs classes est obtenue en combinant un ensemble de classeurs binaires $f_1, f_2, \ldots, f_m$ pour les M classes, et chaque classifier est formé pour différencier une seule classe des autres. La combinaison est effectuée en fonction de la sortie maximale avant l'application de la fonction de signe Weston and Watkins (1998). Maintenant, nous incluons l'attribut démographique dans la sortie d'une classification linéaire de SVM, pour chaque post p_i. Le vecteur ~ vi correspond au contenu et aux attributs démographiques.

$$f(p_i) = \vec{w} \cdot \vec{v_i} + b \qquad (10)$$

2.8. Conclusion

Dans ce chapitre, nous avons sélectionné des formats sémantiques standards tels que RDF et OWL pour représenter, manipuler et accéder à des informations personnelles dans des communautés en ligne. Nous avons présenté le modèle UCC. Ce modèle est construit par la convergence de deux approches (text mining and semantic models) qui s'intéressent à l'analyse et la représentation des données issues des réseaux sociaux. Ce modèle enrichie les profils utilisateur par la catégorisation de contenu partagés et par la déduction et la modélisation des intérêts des utilisateurs. En permettant par la suite le regroupement des utilisateurs selon leurs centres d'intérêts. Nous avons également proposé une nouvelle approche pour améliorer la catégorisation du texte social. Cette approche est basée sur l'intégration des attributs démographiques des utilisateurs dans le processus de la classification.

L'évaluation de notre modèle (le modèle UCC) et de la nouvelle approche de classification proposée est détaillée dans le chapitre suivant.

Chapitre III

Expérimentation et discussion

3. Chapitre 3 : Expérimentation et discussion

3.1. Introduction

Dans cette expérimentation, nous montrerons comment notre modèle peut être utilisé pour fournir une description sémantique et temporelle des données sociales, à travers une application sociale. Cette application est validée en utilisant des ensembles de données issues de Facebook. Premièrement, nous allons construit les modèles d'apprentissage automatique utilisés (l'approche proposée), puis nous les intégrons dans le modèle UCC et nous poursuivant l'expérimentation. Cette dernière va montrer la puissance du modèle UCC.

3.2. Collecte et construction de données d'expérimentation

Les publications (posts) sur Facebook conviennent parfaitement à cette expérimentation. Ceci est en raison de leur abondance et de leur petite taille. De plus, Facebook est un réseau social populaire avec une grande diversité de profils d'utilisateurs, ce qui signifie la disponibilité des informations sociales sur les utilisateurs (attributs démographiques) et du contenu textuel généré à partir de leurs interactions. Ainsi, nous avons collecté un ensemble de données suffisamment volumineux des murs spécifiques, qui sont accessibles au public, aussi, nous avons collecté nos données à l'aide de l'API Facebook. Ces données sont constituées de contenu textuel accompagné des attributs démographiques des utilisateurs.

Nous avons construit deux ensemble de données, le premier ensemble est utiliser pour construire le modèle d'apprentissage, et pour valider également l'approche proposée basée sur l'intégration des attributs démographique dans la classification de texte social. Le deuxième ensemble est utilisé pour la population de l'ontologie UCC afin de valider notre modèle.

- ✓ **Le premier ensemble de données (Dataset-1)**

  Pour assurer la cohérence et la fiabilité des approches proposées, nous avons testé notre classification et nos approches sur un ensemble de 72 900, la moyenne de 243 messages recueillis par utilisateur entre (Juillet et Septembre 2016) à partir de 300 profils Facebook très actifs et réels. Les personnes sont choisies avec soin, ces profils contiennent des informations correctes sur les caractéristiques des utilisateurs. Ces textes ont été annotés manuellement par trois annotateurs humains en huit (8) catégories différentes (art, mode, sport, technologie, affaires, nouvelles, sciences et éducation, et autres). Ou, la catégorie News contient les sous-catégories (politics, crime, entertainment, et autre). Pour des raisons de confidentialité, les noms des utilisateurs sont supprimés et les identifiants Facebook de tous les utilisateurs sont remplacés par des identifiants uniques anonymes.

Nous résumons les statistiques de base de nos données dans les tableaux 3.1. et 3.2.

Catégories	Le nombre de Posts	Le nombre d'utilisateurs
News	8748	127
Business	8311	121
Technologie	13413	133
Art	9112	113
Sport	6561	107
Mode & Fashion	6780	92
Science & Education	8165	56
Autre	11810	263

Tableau 3.1 : Les statistiques de données de dataset-1pour la classification Général.

Les catégories (la classification News)	La distribution (%)
Politics	31.7
Crime	12
Entertainment	34.2
Autre	22.1

Tableau 3.2 :La distribution des catégories pour la classification News (dataset-1).

Connaître des informations telles que le sexe, l'âge et l'état matrimonial des utilisateurs peut aider à détecter leurs intérêts et préférences. Au lieu d'identifier manuellement ces attributs démographiques utiles, nous commençons par les attributs des utilisateurs publics sur Facebook. Nous avons identifié cinq attributs démographiques majeurs : sexe, âge, état matrimonial, travail, et éducation. Pour mesurer quantitativement ces attributs, nous avons résumé leurs détails dans le tableau 3.3.

Attributs	**Valeurs et distribution**
Sexe	femme (42,66%), homme (57,34%)
Age	13-17 ans (19,4%), 18-27 ans (36,8%), 28-37 (27,8%), 38-60 (16%)
Travail	employé (60,3%), sans emploi (20,3%), non concerné (19,4%)
Education	Universitaire (oui, 44,1%), (non, 55,9%)
Etat matrimonial	marié (46,1%), non marié (34,5%), non concerné (19,4%)

Tableau 3.3 : Liste des attributs démographiques.

- ✓ **Le deuxième ensemble de données (Dataset-2)**

  Les données sont collectées du mois d'août 2015 jusqu'au février 2016. Après avoir supprimé les messages non anglais et ceux qui ne contiennent pas d'informations textuelles, toutes les données contenaient 51840 publications postés par 72 utilisateurs (avec ces attributs démographiques), une moyenne de 4 publications par utilisateur par jour. Ces textes ont été annotés manuellement par trois annotateurs humains en huit (8)

catégories différentes (art, mode, sport, technologie, affaires, nouvelles, sciences et éducation, et autres). Ou, la catégorie News contient les sous-catégories (politics, crime, entertainment, et autre). Pour des raisons de confidentialité, les noms des utilisateurs sont supprimés et les identifiants Facebook de tous les utilisateurs sont remplacés par des identifiants uniques anonymes.

Un résumé des distributions des catégories sont présentées dans les Tableaux 3.4. et 3.5.

Les catégories (la classification générale)	La distribution (%)
News	11.5
Business	11.3
Technologie	16.3
Art	14.9
Sport	12.8
Mode & Fashion	11.2
Science & Education	5.3
Autre	10.7

Tableau 3.4: La distribution des catégories pour la classification Générale (dataset-2).

Les catégories (la classification News)	La distribution (%)
Politics	28.9
Crime	14.7
Entertainment	25.6
Autre	30.8

Tableau 3.5 :La distribution des catégories pour la classification News (dataset-2).

3.3. Métriques d'évaluation des modèles d'apprentissage automatique

Nous évaluons maintenant la performance de notre modèle, avant tout, nous devons évaluer les performances des modèles d'apprentissage automatique, F-Measure (F) est l'un des paramètres standard utilisés pour l'évaluation les modèles d'apprentissage automatique, cette mesure comprend deux facteurs fondamentaux, Precision (P) et le Recall (R). Les modèles d'apprentissage sont évalués par : Accuracy (A), Precision (P), le Recall (R) et F-Measure (F). En les obtenant à partir des relations suivantes Lewis (1995) :

$$A = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions}$$

$$p = \frac{Posts\ correctly\ put\ into\ a\ category\ i}{total\ posts \in category\ i}$$

$$R = \frac{Postscorrectlyputintoacategoryi}{Allthepoststhatshouldhavebeen \in thecategoryi}$$

$$F = \frac{2 \times P \times R}{P + R}$$

- **K-fold cross-validation :** nous avons validé nos résultats en utilisant une validation croisée de 10 fold. Comme il n'ya jamais assez de données pour former un modèle d'apprentissage automatique, la suppression d'une partie de celui-ci pour tester pose un problème en réduisant les données d'apprentissage. Par conséquent, nous risquons de perdre des parties importantes d'informations, ce qui augmente l'erreur de classification. Nous avons donc besoin d'une méthode qui fournisse suffisamment de données pour la construction du modèle et laisser également d'autres données pour le test de validation. La validation croisée K-Fold fait exactement cela. Dans la validation croisée K-Fold, les données sont divisées en k sous-ensembles. De sorte que à chaque fois, l'un des k sous-ensembles est utilisé en tant qu'ensemble de test et que les k-1 autres sous-ensembles sont rassemblés pour former un ensemble d'apprentissage. L'estimation de l'erreur est moyennée sur tous les k essais pour obtenir l'efficacité totale de notre modèle. Comme on peut le constater, chaque point de données doit figurer dans un ensemble de validation exactement une fois, et doit se retrouver dans un ensemble d'apprentissage k-1 fois. En outre, l'interchangeabilité des ensembles d'apprentissage et de test renforce également l'efficacité de cette méthode. En règle générale et avec des preuves empiriques, on préfère généralement K = 10 (utilisé pour valider ce travail), mais rien n'est figé et il peut prendre toute valeur.
- **L'outil WEKA: (***Waikato environment for knowledge analysis)*, Witten and Frank (2000) toutes les classifications de ce travail sont construites en utilisant cet outil ou tous les classifiers utilisées sont déjà implémentés.

3.3.1. Métriques d'évacuation de modélisation des intérêts des utilisateurs

Parce que l'un de nos principaux objectifs est de modéliser les intérêts des utilisateurs, d'analyser et de comparer la dynamique des intérêts de l'utilisateur au fil du temps. Nous utilisons la *similarité cosinus* pour évaluer les différences entre les intérêts actuels et les intérêts antérieurs, aussi pour évaluer la similarité entre les intérêts de deux utilisateurs sur une période de temps. Comme nous l'avons décrit dans les sections précédentes, les posts sont classés dans un ensemble prédéfini de catégories dans la CCHT construite.

Dans les analyses des publications, nous avons calculé la répartition des catégories d'intérêt de l'utilisateur sur l'ensemble des catégories pour les utilisateurs individuels ainsi que pour tous les utilisateurs dans l'ensemble de données. Tout d'abord, nous sélectionnons chaque période (par exemple : 20 jours, un mois, deux mois) comme une unité de temps. Ensuite, pour chaque utilisateur, nous allons calculer la répartition de ses intérêts dans chaque période

d, *Du (d)* est représenté comme un vecteur sur l'ensemble des catégories :

$$Du(d) = \left(\frac{s_{c_1}}{S} + \frac{s_{c_2}}{S} + \cdots + \frac{s_{c_m}}{S}\right) \qquad ou \;\; S = \sum s_{u,c_j} \qquad (11)$$

S_{c_j} est le score associé à la catégorie c_j, qui est le nombre de messages classés dans la catégorie c_j publiés par l'utilisateur *u* dans une période donnée *d*. *S* est le nombre total de messages affichés par l'utilisateur dans la période de temps *d*. Ainsi, *Du (d)* représente la proportion de messages publiés par l'utilisateur sur chaque catégorie et reflète également l'importance de chacun des intérêts.

- Similarité d'intérêt

  Nous avons choisi la similarité du cosinus pour comparer la façon dont les intérêts d'un utilisateur changent avec le temps. Compte tenu d'un utilisateur u, qui a le vecteur d'intérêts *Du* (d_1) dans la période de temps d_1 et le vecteur d'intérêts *Du* (d_2) dans la période de temps d_2, où $(d_1 < d_2)$, qui sont représentés par la distribution des catégories d'intérêts en utilisant la même représentation de vecteur que celle utilisée pour un utilisateur donné.

$$Similarity\big(D_u(d_1), D_u(d_2)\big) = \frac{D_u(d_1).D_u(d_2)}{||D_u(d_1)||\;||D_u(d_2)||} \qquad (12)$$

Ici, si la valeur de la similarité est élevée, elle indique que l'utilisateur a des intérêts habituels et réguliers, mais si ses intérêts sont bas, les intérêts du même utilisateur ont un changement important dans le temps. Pour examiner plus avant les performances de ces modèles pour inférer les intérêts des utilisateurs, nous sélectionnons également deux représentants du système de classement des intérêts des utilisateurs et calculons les principaux intérêts pour eux en utilisant les trois modèles.

- Similarité des intérêts entre les utilisateurs

Étant donné les utilisateurs u_1 et u_2 avec les vecteurs d'intérêts $D_{u_1}(d)$ et $D_{u_2}(d)$ à la même période *d* :

$$Similarity\left(D_{u_1}(d), D_{u_2}(d)\right) = \frac{D_{u_1}(d).D_{u_2}(d)}{||D_{u_1}(d)||\;||D_{u_2}(d)||} \qquad (13)$$

Ici, si la valeur de la similarité est élevée, elle indique que les utilisateurs ont de gros intérêts communs dans un certain temps, mais si leur faible niveau, cela signifie que les utilisateurs ont des intérêts très importants. Pour examiner plus avant la performance de ces modèles pour inférer les intérêts des utilisateurs, nous sélectionnons également deux représentants du système de classement des intérêts des utilisateurs et calculons les principaux intérêts pour

eux en utilisant les trois modèles.

Après avoir calculé la similitude entre deux utilisateurs, il est possible d'extraire les intérêts similaires et différents en affichant l'arbre d'intérêt de chaque utilisateur dans la période

3.4. Analyse quantitative du lien entre les attributs démographiques de l'utilisateur et le texte généré par l'utilisateur

Dans les OSN, le texte généré par l'utilisateur contient de riches commentaires, discussions et opinions, ce qui est une ressource importante pour révéler les préférences et les intérêts des utilisateurs. Par conséquent, notre hypothèse est que les attributs démographiques des utilisateurs peuvent être utiles pour prédire leurs intérêts et améliorer les performances de la catégorisation du texte. Il y a eu un manque d'analyse quantitative montrant le lien entre les attributs démographiques d'un utilisateur et le texte généré par l'utilisateur (texte social).

Pour comprendre le lien entre les attributs démographiques des utilisateurs et le texte généré par l'utilisateur, nous proposons de calculer le nombre d'utilisateurs ayant des attributs démographiques similaires dans chaque catégorie. La Figure 3.1 présente les résultats de manière détaillée pour chaque attribut démographique, dans laquelle nous pouvons formuler les observations suivantes: pour comprendre le lien entre les attributs démographiques des utilisateurs et le texte généré par l'utilisateur, nous proposons de calculer le nombre d'utilisateurs ayant des attributs démographiques similaires dans chaque catégorie.

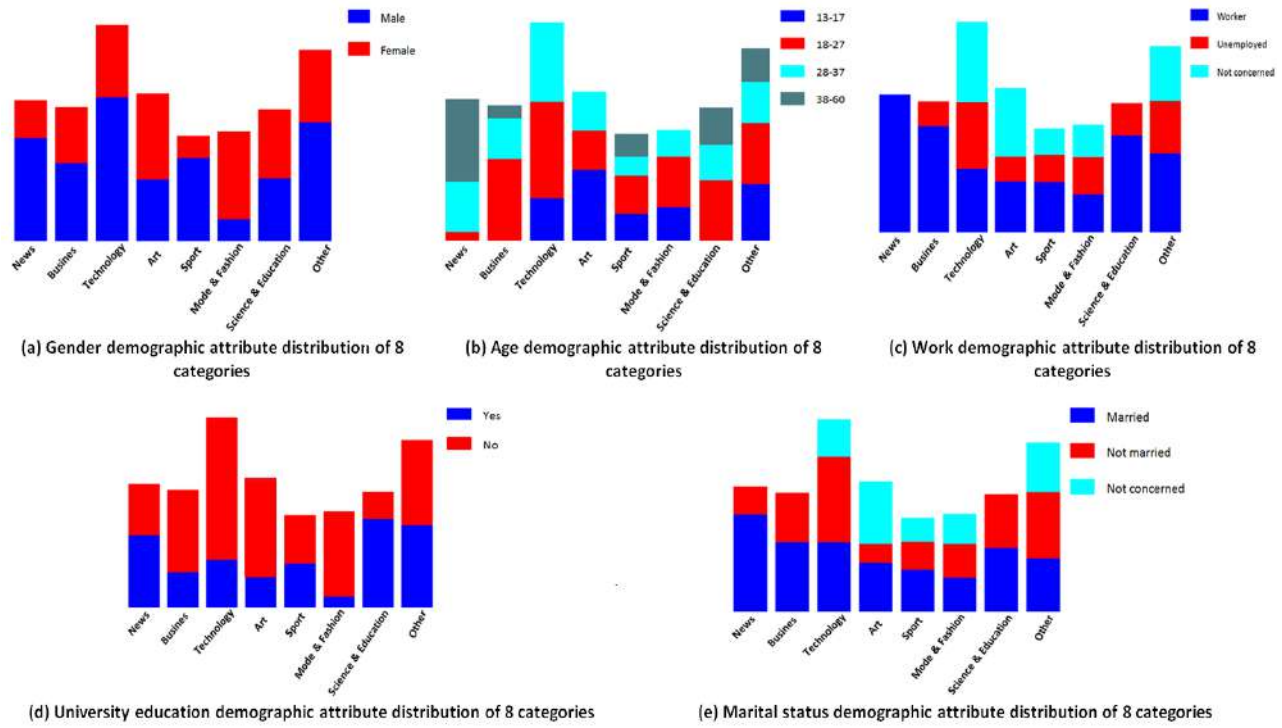

Figure 3.1: les distributions d'attributs démographiques de 8 catégories de la classification générale (dataset-1).

La Figure 3.1 présente les résultats de manière détaillée pour chaque attribut démographique. Nous pouvons formuler les observations suivantes :

- En observant la Figure 3.1., (a), concernant la distribution de l'attribut gender des (8) catégories, nous remarquons que l'homme apparaît davantage dans chacune des catégories (News, Technologie, Sport et Autre). Ou, les femmes apparaissent plus dans les catégories (Art, Mode & Fashion) et presque sont égaux pour le reste des catégories. Nous déduisons que l'attribut gender a un impact fort sur la plupart des catégories (News, Technologie, Sport, Art, Mode & Fashion et Other) et un impact faible sur le reste des catégories.
- La Figure 3.1. (b) représente les distributions de l'attribut Age sur les (8) catégories choisies. Nous remarquons ce qui suit : Les personnes entre (13-17) ont tendance aux catégories (Art et Autres). Ils apparaissent en grande partie sur (Technologie Sport et Mode & Fashion) alors qu'il n'en existe aucune dans chacun des domaines (News, Busines, Science et Education). Les personnes entre (18-27) ans apparaissent dans toutes les catégories dans des proportions variables. Ils couvrent les catégories suivantes : Business, Technologie, Sport, Science et Education et mode & Fashion; de plus, ils apparaissent en grande partie dans chacune des catégories (Art et Other) et avec une très faible proportion dans la catégorie News. Les personnes entre (28-37) apparaissent dans toutes les catégories avec des proportions moyennes, en particulier dans les catégories (Technologie, News). Les gens dans (38-60) couvrent la catégorie (News). Ils apparaissent également dans chacune des catégories (Science e& Education Sport et Other) et avec une très faible proportion dans la catégorie (Business). L'attribut Age affecte toutes les catégories.
- La Figure 3.1., (c) représente la distribution de l'attribut travail (work) sur les (8) catégories choisies. Nous remarquons que les travailleurs couvrent totalement la catégorie News, couvrent presque les catégories (Business, Other et Science & Education) et apparaissent également dans le reste des catégories, tandis que les chômeurs apparaissent dans toutes les catégories sauf News. Les personnes non concernées existent principalement dans les catégories Art et Technologie. Ils apparaissent également dans Sport, Mode & Fashion et Other. L'attribut de travail a principalement une incidence sur Business, News, Sciences, Art.
- La Figure 3.1., (d) représente la distribution de l'attribut éducation sur les (8) catégories choisies. Nous remarquons que les personnes ayant un niveau d'éducation élevé couvrent la plupart des catégories (Science & Education et Actualités), les autres apparaissent davantage dans les catégories (Mode & Mode, Affaires, Technologie et Art) et presque égales pour le reste des catégories. Nous déduisons que, l'attribut de genre à un impact fort sur la plupart des catégories (Actualités, Technologie, Affaires, Art, Science et Education, Mode & Mode et Autres) et un impact faible sur les catégories (Sport et Autres).
- La Figure 3.1., (e) représente la distribution de l'attribut d'état civil sur les (8) catégories choisies. Nous remarquons que les personnes mariées couvrent la plupart des catégories (News et Science & Education). Les personnes non mariées couvrent la plupart de la catégorie (Technologie), les personnes non concernées apparaissant principalement dans (Art) et pas du tout dans (News, Business, et Science & Education). Cet attribut a un fort impact sur les catégories (News, Business,

Science & Education et Art) et faible sur les catégories (Technologie, Sport, Mode & Fashion et Other).

En général, nous déduisons que le lien entre les attributs démographiques des utilisateurs et le texte généré par l'utilisateur est fort ou faible par rapport à la catégorie en question. En d'autres termes, chaque attribut a un impact important sur certaines catégories et un faible impact sur autres, comme indiqué ci-dessus.

3.5. Prétraitement de texte et les « term weighting schemes » utilisés

- ✓ Un terme qui apparaît moins de trois (3) fois est supprimé ;
- ✓ Suppression de la ponctuation (.,!?) Et des symboles ([<>)) ;
- ✓ Le stemmer utilisé est lovenStemmer qui est largement utilisé dans la littérature.

Dans ce travail, quatre différents term weighting schemes supervisés et non supervisés sont choisis (Binary, TFIDF, Chis, OddsR). Ces derniers sont choisis en raison de leurs résultats supérieurs rapportés et de leur représentation typique dans la catégorisation du texte.

3.6. Résultats et discussion

3.6.1. Résultat des classifications

Dans cette section, nous évaluons la performance des approches proposées. Maintenant, notre expérimentons le modèle d'apprentissage en utilisant la validation croisée par k=10 avec WEKA Witten and Frank (2000), où tous les classifiers utilisés sont déjà implémentés. Les Figures de 3.2 au 3.7 illustrent les résultats de F-mesure pour chaque classifier utilisée dans chaque version de données (V1 à V7), en utilisant respectivement les TWS (Binary, TFIDF, ChiS et OddsR) comme modèles de sélection des caractéristiques (feature selection).

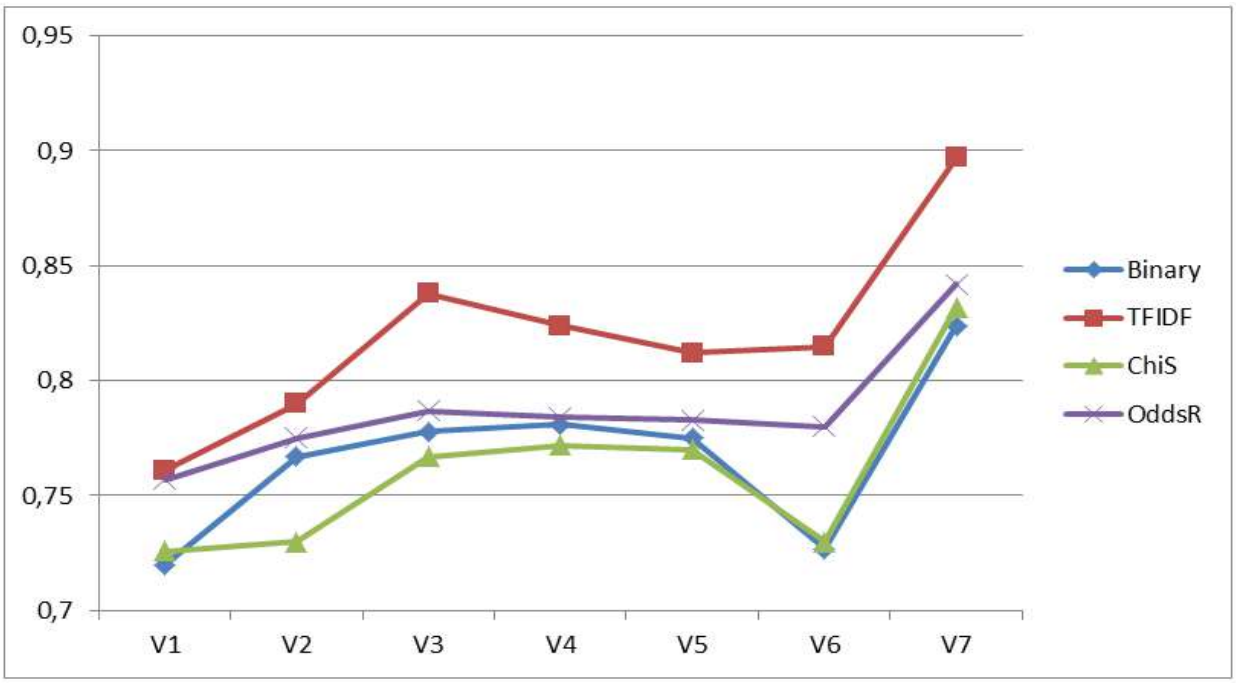

Figure 3.2: les résultats de F-mesure obtenus en utilisant le classifier ANN sur les versions de données (V1 à V7).

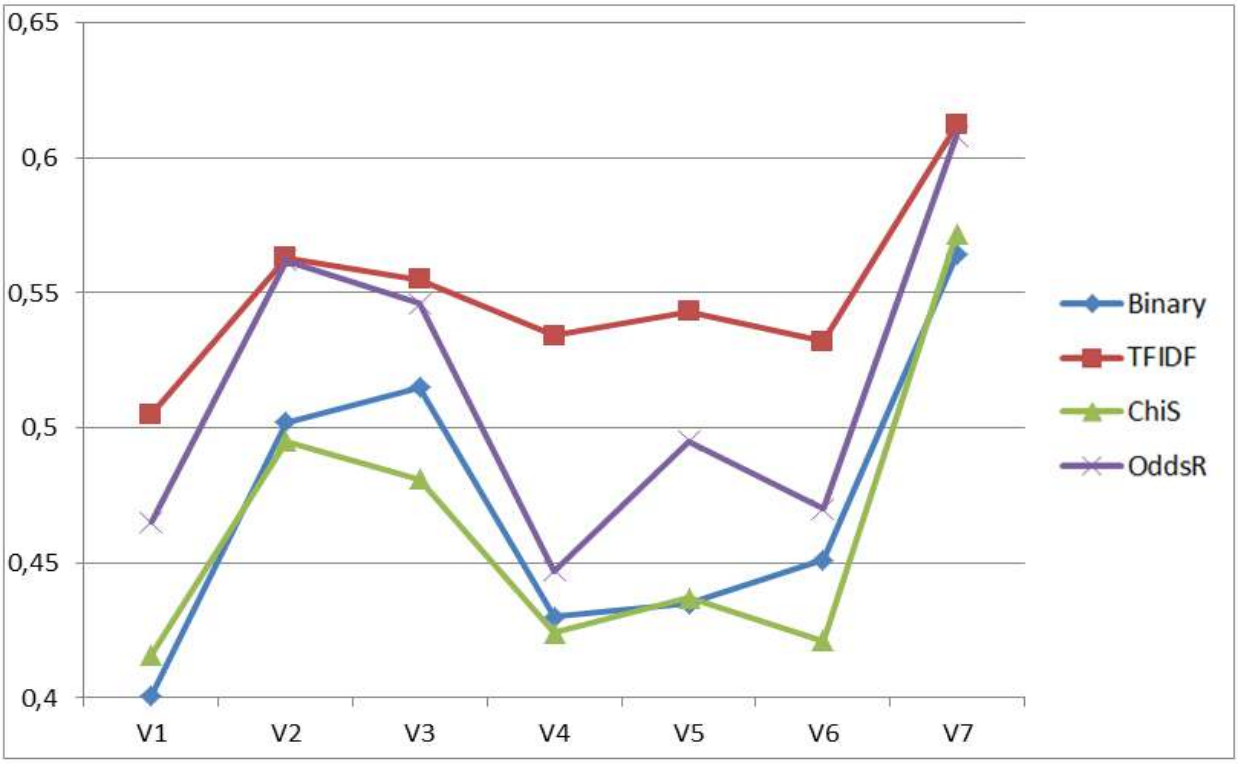

Figure 3.3: les résultats de F-mesure obtenus en utilisant le classifier KNN sur les versions de données (V1 à V7).

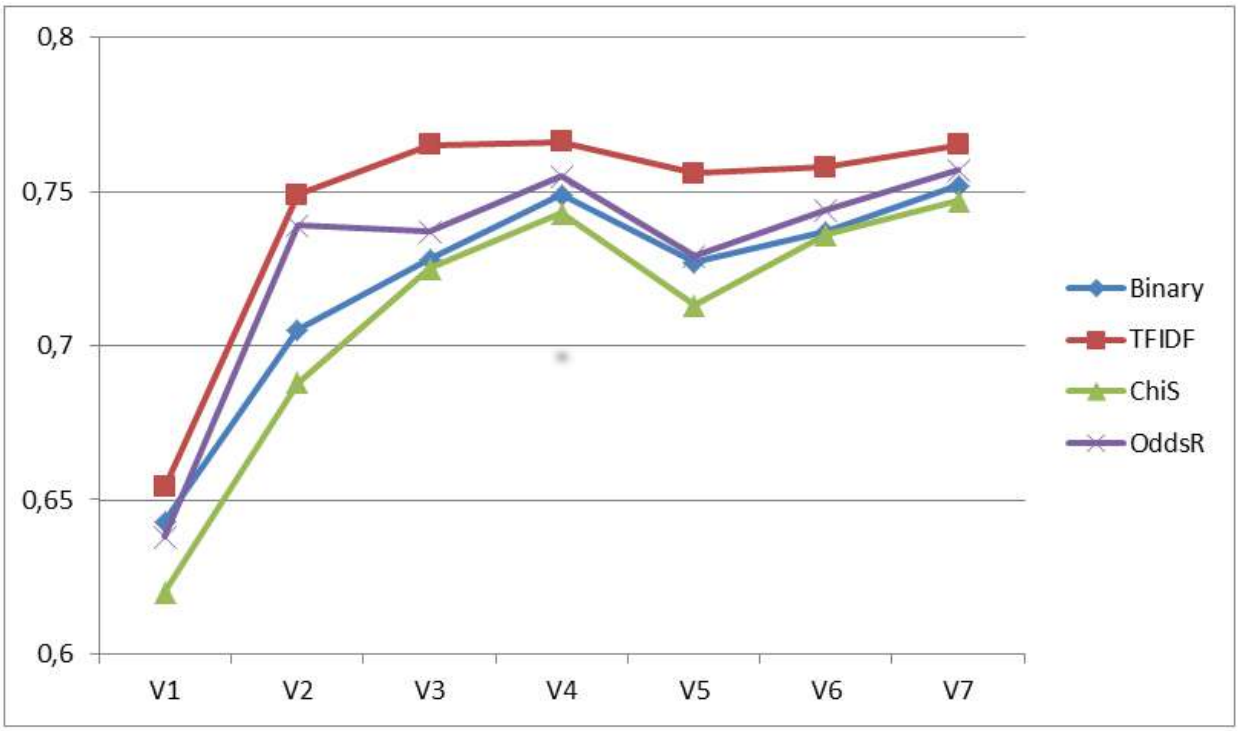

Figure 3.4: les résultats de F-mesure obtenus en utilisant le classifier NB sur les versions de données (V1 à V7).

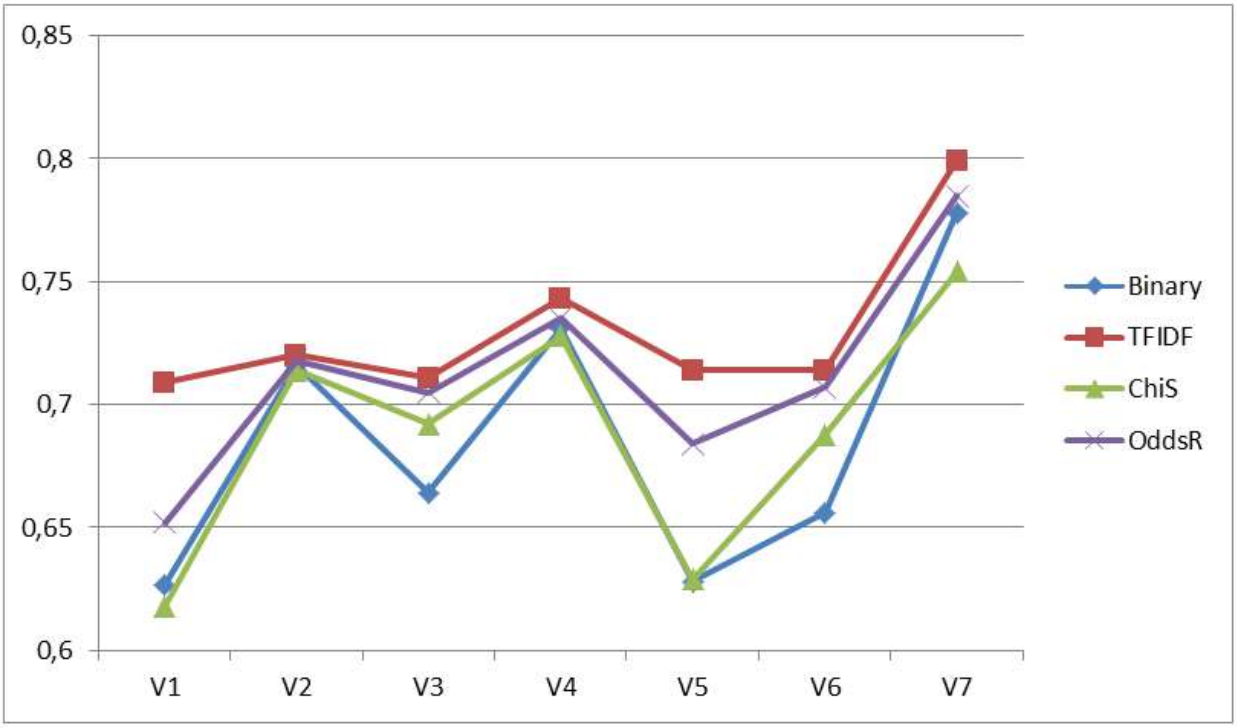

Figure 3.5: les résultats de F-mesure obtenus en utilisant le classifier DT sur les versions de données (V1 à V7).

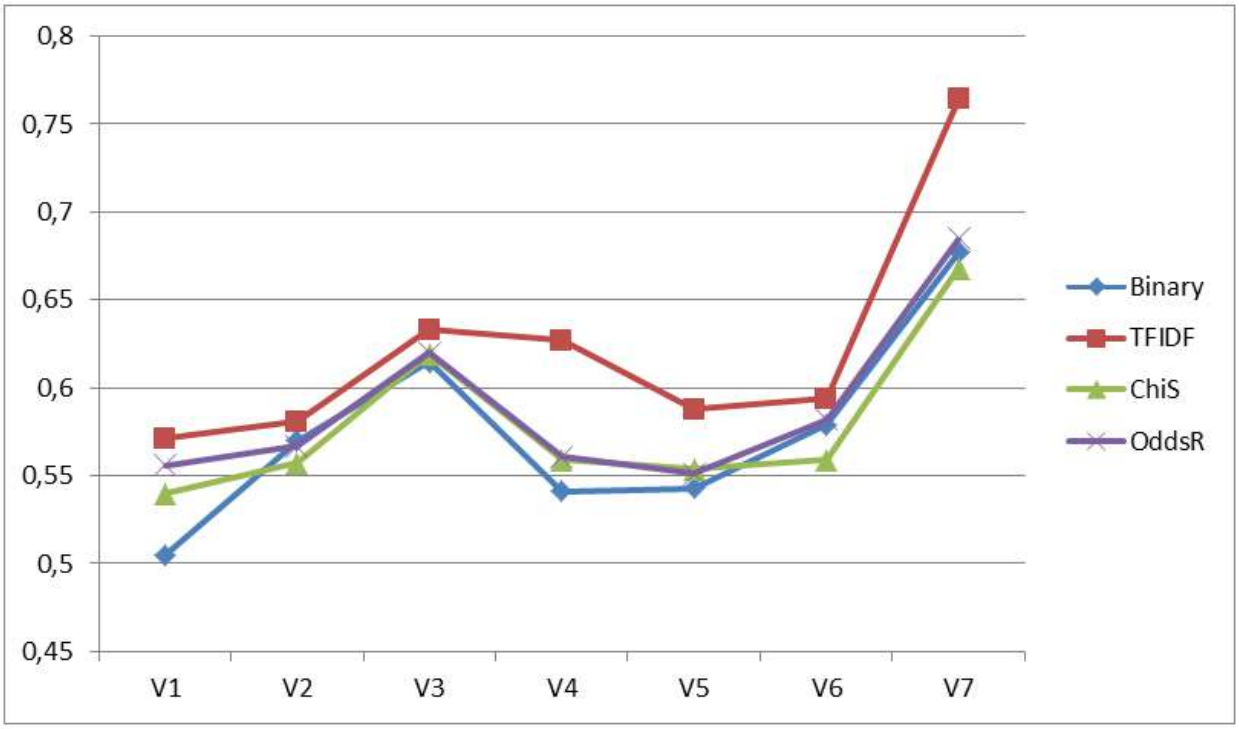

Figure 3.6: les résultats de F-mesure obtenus en utilisant le classifier DR sur les versions de données (V1 à V7).

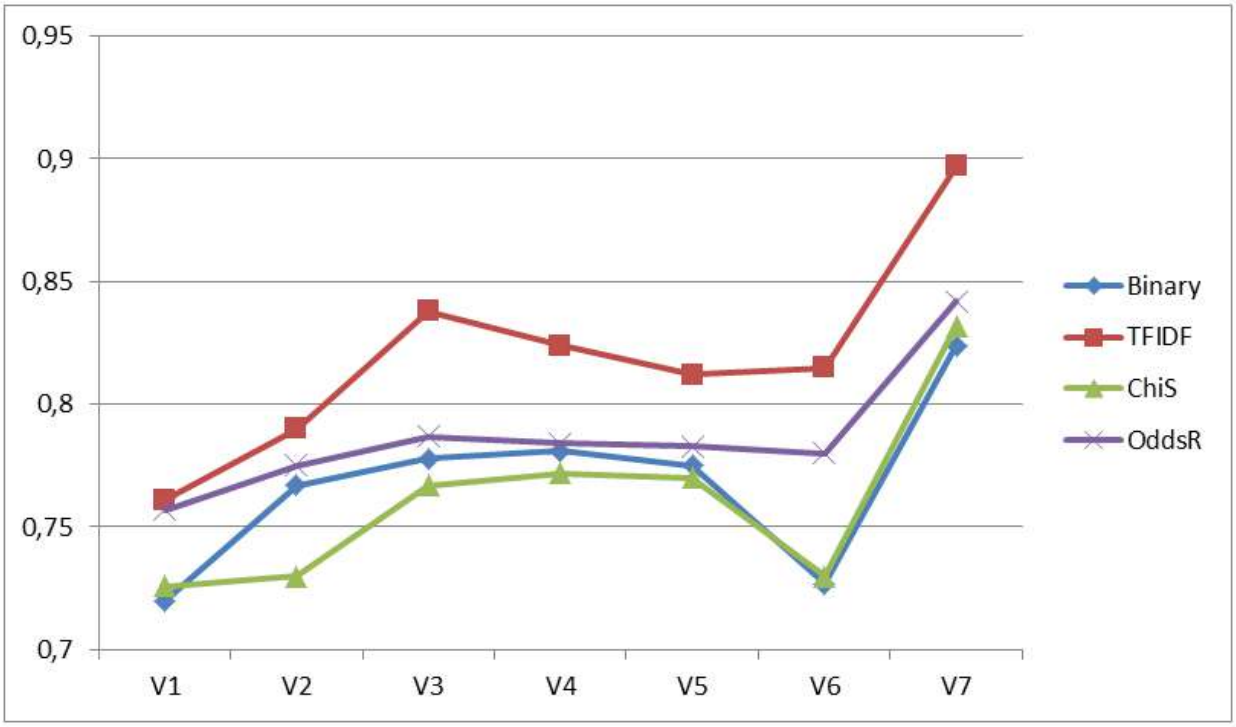

Figure 3.7: les résultats de F-mesure obtenus en utilisant le classifier SVM sur les versions de données (V1 à V7).

✓ **Une comparaison entre les résultats obtenus des TWS pour les six (6) classifiers utilisés pour toutes les versions de données (V1 à V7)**

Selon les Figures du 3.2. au 3.7., en comparant les résultats obtenus en utilisant les versions de données (V2 à V7) avec les résultats obtenus en utilisant (V1) concernant tous les TWS utilisés, nous notons une amélioration remarquable des résultats des F-mesures pour tous les classifiers.

- De la Figure 3.2., en utilisant le classifier ANN, pour toutes les versions de données (V1 à V7), le modèle basé sur TFIDF a surperformé les autres. Suivi par le modèle basé sur OddsR, puis sur le modèle basé sur binary, tandis que le modèle basé sur le ChiS a obtenu les plus mauvais résultats.
- La Figure 3.3. représente les résultats de F-mesure obtenus en utilisant le classifier KNN. Pour toutes les versions de jeu de données (V1 à V7), nous remarquons que le modèle basé sur TFIDF a obtenu les meilleurs résultats et que le modèle basé sur OddsR s'est classé deuxième. Le modèle basé sur le Binary fonctionne mieux que le modèle basé sur le Chis en V2, V3, V4 et V6, alors que dans le modèle V1, V5 et V7, le modèle basé sur ChiS a obtenu de meilleurs résultats que le modèle basé sur le binaire.
- La Figure 3.4. représente les résultats de F-Measure obtenus par NB. Pour toutes les versions de jeu de données (V1 à V7), le modèle basé sur TFIDF a surperformé les autres. Suivi par le modèle basé sur OddsR, puis sur le modèle basé sur le Binary, tandis que le modèle basé sur le ChiS a obtenu les plus mauvais résultats.
- Sur la Figure 3.5., en utilisant la classifier DT, pour toutes les versions de jeu de données (V1 à V7), nous ne pensons pas que le modèle basé sur TFIDF a obtenu les meilleurs résultats et que le modèle basé sur OddsR se classe au deuxième rang. Le modèle Binary fonctionne mieux que les

modèles basés sur Chis en V1, V2, V4 et V7. En V3, V5 et V6, le modèle basé sur ChiS a obtenu de meilleurs résultats que le modèle basé sur Binary.

- La Figure 3.6. montre les résultats F-measure obtenus à l'aide du classeur DR. Le modèle basé sur TFIDF a obtenu le meilleur résultat pour toutes les versions du jeu de données. Le modèle basé sur OddsR a obtenu de meilleurs résultats que les modèles binaire et ChiS pour toutes les versions de jeu de données, à l'exception de V2, obtenu par un modèle basé sur binaire. Ce dernier a obtenu de meilleurs résultats que ChiS dans les versions V5, V6 et V7, où le modèle basé sur ChiS a obtenu de meilleurs résultats que le modèle basé sur Binary dans le reste des versions de jeu de données.
- La Figure 3.7. montre les résultats de F-measure obtenus utilisant SVM. Le modèle basé sur TFIDF a obtenu le meilleur résultat pour toutes les versions du jeu de données (V1 à V7), suivi des modèles basés sur OddsR. Puis le modèle Binary pour V2, V3, V4 et V5 et modèle basé sur Chis pour les versions de jeux de données V1, V6 et V7.

Pour les six (6) classifies utilisées, il est évident que les modèles basés sur TFIDF obtiennent les résultats les plus élevées de F-Measure. Suivi par les modèles basés sur OddsR, tandis que les performances des résultats de la F-mesure sont alternativement liées d'un classifier à une autre et d'une version de données à une autre entre les modèles basés sur Binary et les modèles basés sur ChiS.

Étant donné que TFIDF fonctionne mieux que les autres TWS dans tous les classifiers et dans toutes les versions de données, nous le choisissons comme représentant des caractéristiques textuelles pour le reste de nos expérimentations. Le tableau 3.6 montre en détail les résultats, y compris Accuracy (A), Precision, Recall and F-mesure, obtenus par les six classifiers dans chaque version de données de (V1 à V7) en utilisant TFIDF.

Dataset	classifier	A(%)	P	R	F
V1 (baseline)	ANN	74.6	0.796	0.748	0.746
	KNN	51.9	0.628	0.519	0.505
	NB	65.2	0.656	0.652	0.654
	DT	71.2	0.714	0.712	0.709
	DR	56.9	0.677	0.569	0.571
	SVM	75.55	0.774	0.755	0.761
V2	ANN	75.9	0.797	0.759	0.757
	KNN	56.6	0.663	0.566	0.563
	NB	74.9	0.754	0.749	0.749
	DT	72.3	0.728	0.723	0.72
	DR	58.2	0.693	0.582	0.581
	SVM	79	0.793	0.79	0.79
V3	ANN	77.9	0.809	0.779	0.778

	KNN	56.6	0.663	0.566	0.563
	NB	74.9	0.754	0.749	0.749
	DT	71.6	0.719	0.716	0.711
	DR	62.7	0.703	0.627	0.633
	SVM	83.9	0.84	0.839	0.838
V4	ANN	80.2	0.806	0.802	0.801
	KNN	53.5	0.665	0.535	0.534
	NB	76.6	0.773	0.766	0.766
	DT	74.6	0.747	0.746	0.743
	DR	62.5	0.704	0.625	0.627
	SVM	82.5	0.825	0.825	0.824
V5	ANN	79.6	0.801	0.796	0.795
	KNN	54.2	0.676	0.542	0.543
	NB	75.4	0.765	0.754	0.756
	DT	71.9	0.721	0.719	0.714
	DR	59.1	0.695	0.591	0.588
	SVM	81.4	0.814	0.814	0.812
V6	ANN	79.3	0.798	0.793	0.793
	KNN	53.3	0.659	0.533	0.532
	NB	75.8	0.765	0.758	0.758
	DT	71.18	0.718	0.718	0.714
	DR	58.9	0.708	0.589	0.594
	SVM	81.6	0.817	0.816	0.815
V7	ANN	87.6	0.876	0.876	0.875
	KNN	68.8	0.62	0.688	0.612
	NB	76.4	0.772	0.764	0.765
	DT	80.2	0.8	0.802	0.799
	DR	76.5	0.783	0.765	0.764
	SVM	90.01	0.896	0.9	0.897

Tableau 3.6:les attributs démographique de l'utilisateur sélectionné dans chaque version de jeu de données (V1 à V7), Precision (P), Recall (R), F-Measure (F), and Accuracy (A), utilisant la validation croisée de 10 fold.

- ✓ **Comparaison entre les résultats des classifiers dans chaque version de données**

Les résultats les plus élevés que nous avons obtenus sont les suivants : accuracy de 90,01%, précision de 0,896, recall de 0,96 et F-Measure en utilisant le jeu de données V7, y compris toutes les caractéristiques des utilisateurs sélectionnés par le classifieur SVM. Premièrement, en utilisant le jeu de données V1 (baseline), nous nous sommes concentrés

uniquement sur le contenu textuel partagé sur Facebook sans tenir compte des attributs démographiques. Nous avons obtenu des accuracy de ces classeurs de 51,9% avec k-NN, 58,2% avec DT, 56,9% avec NB, 71,6% avec DT, 74,6% avec ANN et 75,55%, ce qui représente la meilleure précision obtenue par le SVM. En terme de précision, ANN a obtenu le meilleur résultat avec 0,796 suivi de SVM avec 0,774, DT avec 0,719, DR avec 0,677, NB avec 0,656 et k-NN avec 0,628. En termes de recall et de F-Measure, nous avons obtenu ces résultats, SVM avec (0,755 et 0,761) représente les meilleurs résultats, suivi de ANN avec (0,748 et 0,746), DT avec (0,716 et 0,711), NB avec (0,652 et 0,654), DR avec (0,569 et 0,571) et finally k-NN avec (0,519 et 0,505).

Maintenant, nous montrons les résultats obtenus, y compris les caractéristiques des utilisateurs (attributs démographiques des utilisateurs) dans le processus de classification, où, dans chaque version des données (V2 à V6), nous incorporons l'un de ces attributs. Dans la version de données V7, nous avons inclus tous les attributs. Tous les résultats sont présentés dans le Tableau 3.7.

Dans la version V2 des données, nous avons inclus le genre de l'auteur. Ainsi, SVM a obtenu les meilleurs résultats avec (79%, 0,793, 0,79 et 0,79) en termes d'accuracy, de précision, de recall et de F-Measure, suivis de ANN avec (75,9%, 0,797, 0,759 et 0,757), NB avec (74,9%, 0,754, 0,749 et 0,749), DT avec (72,3%, 0,728, 0,723 et 0,72), DR (58,2%, 0,693, 0,582, 0,581) et finalement kNN avec (56,6%, 0,663, 0,566 et 0,563).

Dans la version V3 des données, nous avons inclus l'attribut age, dans lequel nous nous sommes concentrés sur quatre intervalles différents de l'âge. Les meilleurs résultats sont obtenus par SVM (83,9%, 0,84, 0,839 et 0,838) en accuracy, précision, recall et F-Measure, suivis de ANN avec (77,9%, 0,809, 0,779 et 0,778), NB avec (76,5%). %, 0,773, 0,765 et 0,765), DT avec (71,2%, 0,714, 0,712 et 0,709), DR avec (62,7%, 0,703, 0,627 et 0,633) et k-NN avec (56,1%, 0,67, 0,561 et 0,555).

Dans la V4, nous avons inclus l'aspect du travail (si l'utilisateur est un travailleur ou un chômeur). Les meilleurs résultats sont obtenus par SVM (82,5%, 0,825, 0,825 et 0,824) en accuracy, précision, recall et F-Measure, suivis du ANN avec (80,2%, 0,806, 0,802 et 0,801), NB avec (76,6%). %, 0,773, 0,766 et 0,766), DT avec (74,6%, 0,747, 0,746, 0,743), DR avec (62,5%, 0,704, 0,625 et 0,627) et k-NN avec (53,5%, 0,665, 0,535 et 0,534).

L'éducation de l'utilisateur est concédée dans la V5 (si l'utilisateur a un niveau d'éducation élevé), où les meilleurs résultats sont obtenus par SVM (81,4%, 0,814, 0,814 et 0,812) en termes d'accuracy, précision, recall et F-Measure, ANN, avec (79,6%, 0,801, 0,796 et

0,795), NB avec (75,4%, 0,765, 0,754 et 0,756), DT avec (71,9%, 0,721, 0,719 et 0,714), DR avec (59,1%, 0,695, 0,591 et 0,588) et k-NN avec (54,2%, 0,676, 0,542 et 0,543).
La dernière caractéristique concédée par ce travail est l'état matrimonial (marié ou non marié) inclus dans la V6, où les meilleurs résultats sont obtenus par SVM (81,6%, 0,817, 0,816 et 0,815) en termes accuracy, précision, recall et F-Measure, suivis de ANN avec (79,3%, 0,798, 0,793 et 0,793), NB avec (75,8%, 0,765, 0,758 et 0,758), DT avec 545 (71,18%, 0,718, 0,718 et 0,714), DR avec (58,9%, 0,708, 0,589 et 0,594) et k-NN avec (53,3%, 0,659, 0,533 et 0,532).

Enfin, nous avons rassemblé toutes ces caractéristiques de l'auteur du post dans le jeu de données V7, où SVM surpasse les autres classifiers avec (90,01%, 0,896, 0,9 et 0,897) respectivement en accuracy, précision, recall et F-Measure, suivis de ANN. Avec (87,6%, 0,876, 0,876 et 0,875), DT avec (80,2%, 0,8, 0,802 et 0,799), DR avec (76,5%, 0,783, 0,765 et 0,764), NB avec (76,4%, 0,772, 0,764). et 0,765) et k-NN avec (68,8%, 0,062, 0,688 et 0,612).

- ✓ **Comparaison entre la version de base et les autres versions de données**

En comparant les résultats obtenus avec les versions de données (V2-V7) aux résultats de base (V1), nous notons une amélioration remarquable des mesures d'accuracy, précision, recall et F-Measure, pour tous les classifiers.
En utilisant V2, les accuracie sont améliorées (de 74,6% à 75,9% en utilisant ANN) (de 51,9% à 56,6% en utilisant k-NN) (de 65,2% à 74,9% en utilisant NB) (de 71,6% à 72,3% en utilisant DT).), (de 56,9% à 58,2% avec DR) et pour SVM (de 75,55% à 81,7%).
Pour V3, le résultat de la performance de tous les classifiers par rapport à la baseline (V1) est amélioré (de 74,6% à 77,9% avec ANN) (de 51,9% à 56,1% avec k-NN), (de 65,2% à 76,5%). % utilisant NB) (de 71,6% à 71,2% utilisant DT) (de 56,9% à 62,7% utilisant DR) et pour la classe SVM (de 75,55% à 83,3%).
En utilisant le jeu de données V4, les performances de tous les classifiers sont améliorées (de 74,6% à 80,2 avec ANN) (de 51,9% à 53,5% avec k-NN), (de 65,2% à 76,6% avec NB) (de 71,6% à 74,6 en utilisant DT), (de 56,9% à 62,5% en utilisant DR) et pour la classe SVM (de 75,55% à 82,1%).
En utilisant V5, nous observons une amélioration de la performance de tous les classifiers (de 74,6% à 79,6 en utilisant ANN) (de 51,9% à 54,2% en utilisant k-NN), (de 65,2% à

75,4% en utilisant NB) (de 71,6% à 71,9% en utilisant DT), (de 56,9% à 59,1% en utilisant DR) et pour la classe SVM (de 75,55% à 81,4%).

En utilisant la V6, les performances de la classe sont augmentées (de 74,6% à 79,3% en utilisant ANN) (de 51,9% à 53,3% en utilisant k-NN), (de 65,2% à 75,8% en utilisant NB) 585 (de 71,6% à 71,18% en utilisant DT), (de 56,9% à 58,9% en utilisant DR) et (de 75,55% à 81,2% en utilisant SVM).

Enfin, nous avons rassemblé toutes ces caractéristiques de l'auteur du post dans la série de données V7, où nous avons obtenu les meilleurs résultats pour tous les classifiers (de 74,6% avec 87,6% ANN) (de 51,9% à 68,8% avec k-NN), (de 65,2% à 76,4% en utilisant NB), (de 71,6% à 80,2% en utilisant DT), (de 56,9% à 76,5% en utilisant DR) et (de 75,55% à 90,01% en utilisant SVM) dans les accuracies des classifiers.

- ✓ **Comparaison des résultats des classeurs à l'aide de la version de données V7**

Comme mentionné précédemment, en utilisant la version V7 des données (y compris tous les attributs démographiques), nous avons obtenu les meilleurs résultats pour tous les classifiers, le plus élevé étant obtenu en utilisant SVM avec une accuracy de 90,01%, une précision de 0,896, un recall de 0,9 et 0,96 en F-Measure, suivi du classificateur ANN avec une accuracy de 87,6%, une précision de 0,876 en recall, et de 0,875 en F-Measure, puis du classeur DT avec une accuracy de 80,2%, une précision de 0,8, un recall de 0,802 et de 0,799 en F-Measure, suivi de DR avec une Accuracy de 76,5%, une précision de 0,783, un recall de 0,765 et une F-Measure de 0,764, ensuite, NB avec une précision de 76,4%, une précision de 0,772, un rappel de 0,764 et de 0,765 en F -mesure, où k-NN avec une accuracy de 68,8%, une précision de 0,062, un recall de 0,688 et une F-Measure de 0,612 a obtenu le résultat le plus faible.

Les performances remarquables de SVM montrent que SVM a toujours de meilleures performances en termes de classification de texte, ce qui fait référence à sa capacité à classer les données comportant des caractéristiques de grande dimension. Cela correspond à la caractéristique de SVM qui est capable de généraliser bien dans les caractéristiques de grande dimension et omet également la nécessité de la sélection de caractéristique.

- ✓ **Comparaison des résultats SVM avec V7 et V1 (base)**

Le Tableau 3.7 montre une comparaison entre les meilleurs résultats obtenus avec (V7) et les résultats de base obtenus avec (V1) par SVM pour chaque catégorie.

	V1				V7			
Catégories	A(%)	P	R	F	A(%)	P	R	F
News	69.7	0.762	0.697	0.728	84.6	0.835	0.846	0.84
Business	72.9	0.927	0.729	0.816	82.4	0.824	0.824	0.824
Technology	96.8	0.945	0.968	0.956	96.2	0.959	0.962	0.96
Art	74	0.77	0.74	0.755	95.2	0.969	0.952	0.961
Sport	58.4	0.577	0.584	0.581	94.2	0.953	0.942	0.947
Mode & Fashion	83.7	0.862	0.837	0.85	87	0.957	0.87	0.911
Science & Education	68.1	0.75	0.681	0.714	89.3	0.834	0.893	0.863
Other	89	0.689	0.89	0.777	91.2	0.844	0.912	0.877

Tableau 3.7: Les résultats obtenus par SVM avec les versions de jeu de données V7 et V1 (baseline) pour chaque catégorie.

En analysant les résultats obtenus par le classifier SVM, nous remarquons que l'inclusion des attributs démographiques a un impact positif sur les performances du classifier, qui est améliorée de 14,46%. Cette approche améliore les résultats de l'accuracy pour toutes les catégories : news de 14,9%, business de 9,5%, art de 21,2%, sport de 35,8%, 3,3% dans la catégorie mode & fashion, 21,2% dans science & éducation, et 2,2% dans la catégorie Other, où nous remarquons une détérioration de 0,6% de l'accuracy de la categorie technologie.

Les résultats montrent la force de la prise en compte des attributs démographiques dans la performance de la classification du texte social. Tous les classeurs utilisés ont montré une amélioration des performances de classification en considérant un attribut démographique dans chaque version de jeu de données (V2 à V6) et tous (V7) par rapport au V1 (en considérant uniquement les caractéristiques textuelles). Ces résultats correspondent aux résultats de l'analyse quantitative, qui montrent l'existence d'un lien entre les attributs démographiques des utilisateurs et le texte généré par l'utilisateur. Ce lien est fort ou faible relativement à la catégorie d'intérêt ; ce résultat a été confirmé par l'observation des résultats de SVM pour toutes les catégories, où chaque attribut démographique a un impact sur des catégories d'intérêts spécifiques.

Après la construction de modèle d'apprentissage automatique pour la classification générale, maintenant, nous allons construire un modèle l'apprentissage automatique pour la classification News.

	A	P	R	F
SVM-TFIDF	71.9	0.728	0.719	0.719

Tableau 3.8 : Les résultat d'apprentissage utilisant SVM pour la classification News

Après avoir construit les deux modèles d'apprentissage automatique utilisé dans notre application, nous construisant notre CCHT, représentée dans la Figure 3.9.

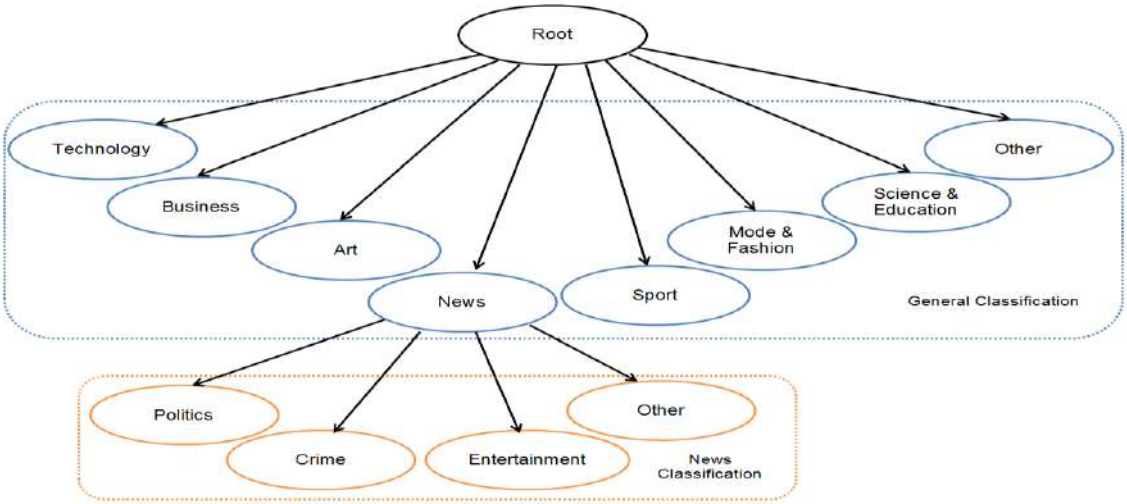

Figure 3.8: la CCHT construite de l'apprentissage pour l'application.

3.6.2. Analyser la dynamique des intérêts des utilisateurs

Dans cette section, nous nous intéressons à la construction d'UIT, à savoir l'arbre d'intérêts des utilisateurs. Pour accomplir cette tâche pour chaque utilisateur, nous avons comparé la répartition de ses intérêts du mois le plus récent à la répartition de ses intérêts de tous les mois précédents (1er septembre 2015 au 29 février 2016). En prenant l'exemple de l'utilisateur User_A, la Figure 3.9. Représente la similarité entre les intérêts de l'utilisateur User_A en février (le mois en cours) et ceux des mois précédents.

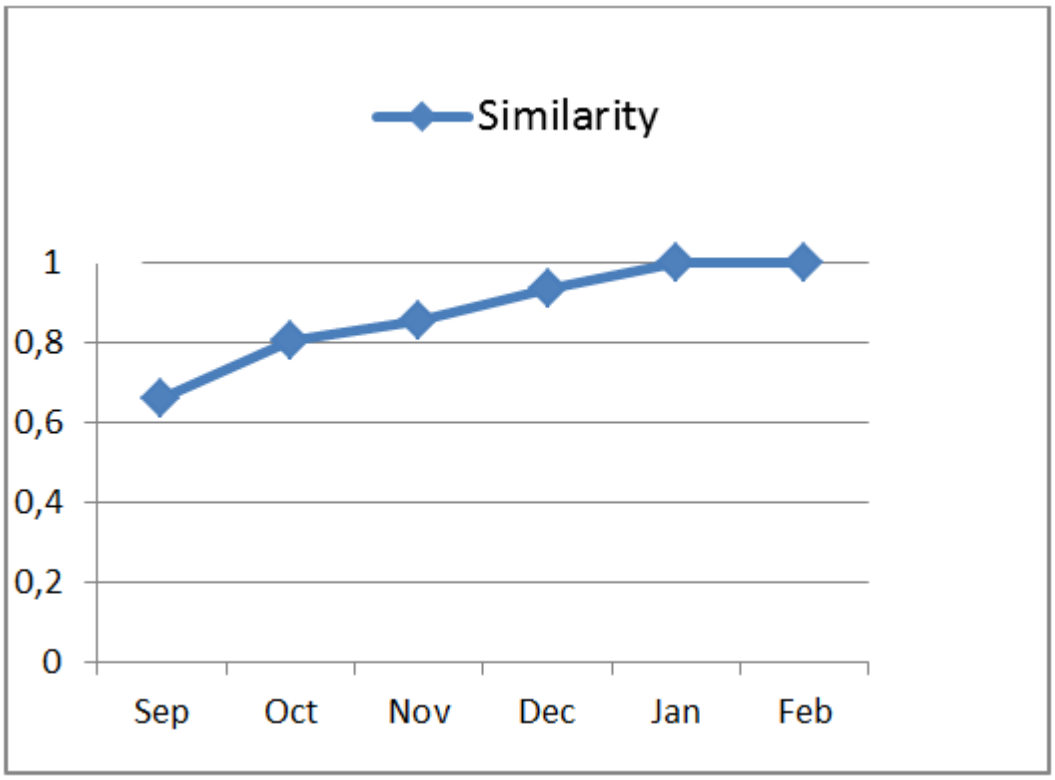

Figure 3.9 : la similarité des intérêts de l'utilisateur User_A entre le mois en cours et les mois précédents.

En observant la similarité des intérêts actuels avec les intérêts précédents générés au début, nous en déduisons qu'il s'agit du minimum, alors que la similarité des intérêts actuels avec les intérêts calculés le mois dernier est la plus élevée. Cette comparaison permet de créer de nouveaux profils d'intérêt qui répondent mieux aux demandes actuelles des utilisateurs.

En raison de l'espace limité, nous ne montrons que l'évaluation des 5 principales catégories d'intérêts (les catégories représentées par les feuilles de l'UIT) de l'utilisateur User-A

Les autres utilisateurs ont un cas similaire. Les résultats sont illustrés à la Figure 3.10.

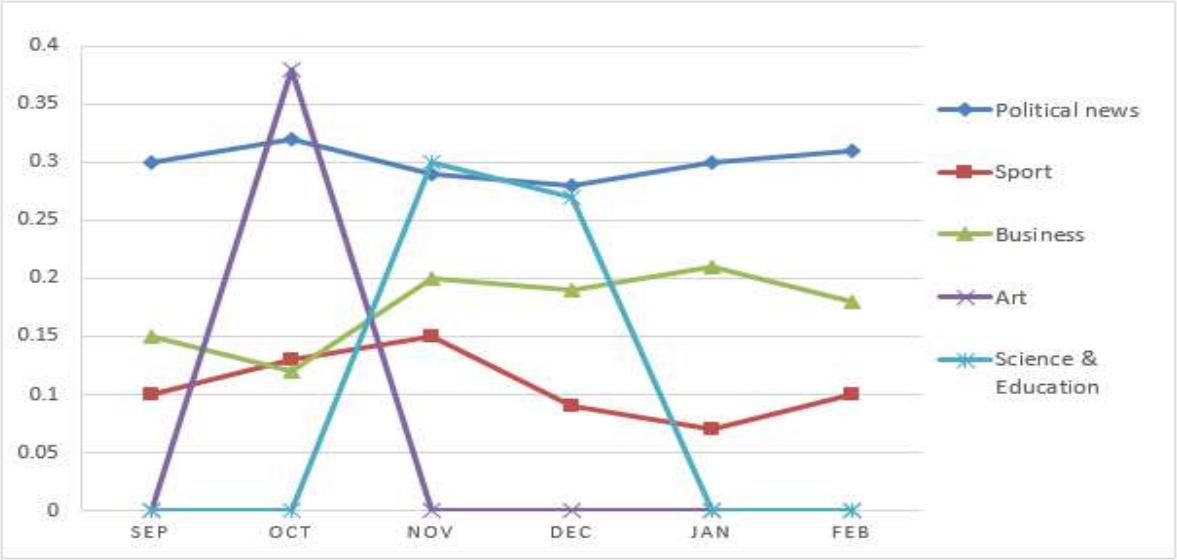

Figure 3.10: évolution périodique des 5 principales catégories d'intérêts de l'utilisateur User-A

Comme nous l'avons mentionné précédemment, les intérêts de l'utilisateur peuvent être divisés en un intérêt permanent, ce qui est une préférence inhérente originale, où il maintient la stabilité dans une période de long terme (dans le cas de l'utilisateur User-A, les catégories d'intérêts permanents sont Political News, Sport et Business). Où, les intérêts temporaires : lorsque l'utilisateur suit les pistes de sujets chauds ou d'événements dans le moment. Par exemple, l'utilisateur User-A s'intéresse à la catégorie Art pour une courte période.

3.6.3. La similarité d'intérêt entre les utilisateurs

Dans cette section, nous sommes intéressés par les utilisateurs qui ont des intérêts approximatifs et les utilisateurs intéressés par la même catégorie d'intérêt. Pour accomplir cette tâche, pour chaque utilisateur, nous calculons la similarité entre les distributions de chaque utilisateur de deux utilisateurs dans la même période. La Figure 3.11. représente la similarité des intérêts de Liam avec 3 autres utilisateurs (User-A, User-B et User-C) dans la période de 4 mois**.**

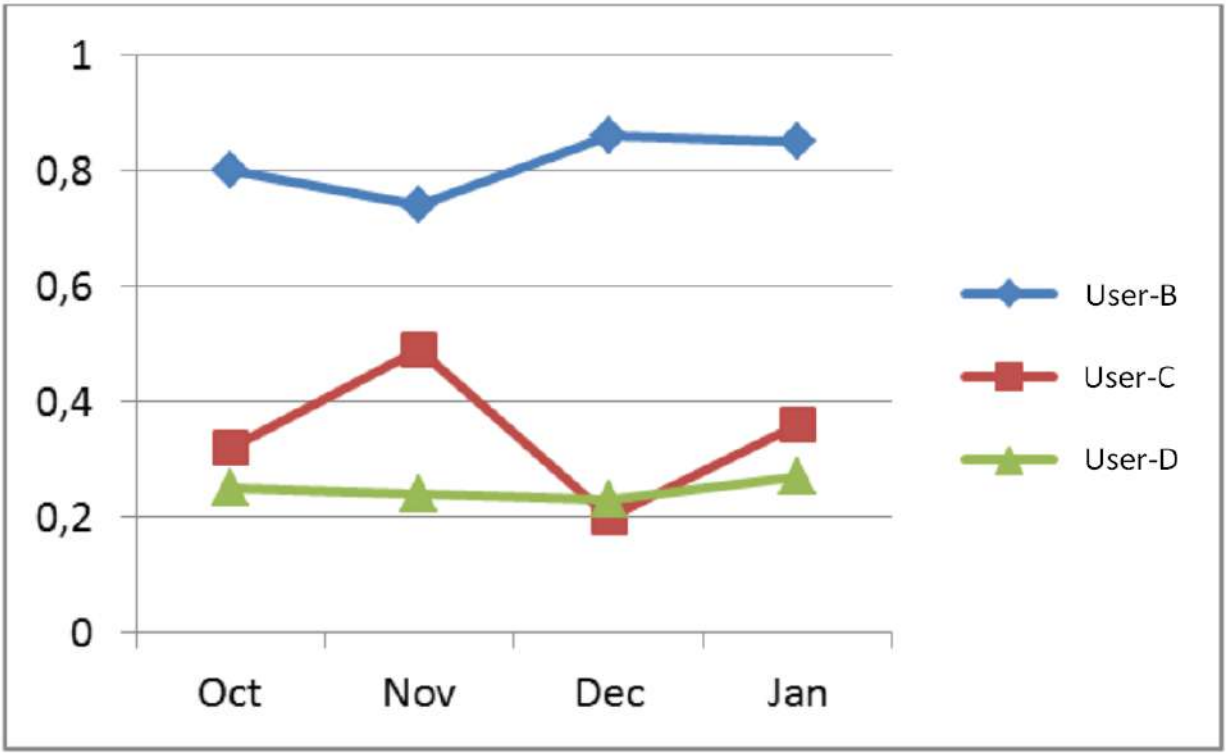

Figure 3.11: la similarité entre l'utilisateur User-A et 3 autres utilisateurs

Nous constatons que les intérêts de l'utilisateur User-B sont les plus semblables aux intérêts de l'utilisateur User-A. Et nous pouvons conclure que les utilisateurs User-B et User-A ont une plus grande partie des intérêts communs. De la même manière, nous pouvons comparer tous les intérêts des utilisateurs avec les intérêts des utilisateurs de nos données aux autres et les regrouper.

3.6.4. Recherche des données par l'ontologie UCC

Dans cette section, nous utilisons SQWRL pour interroger les données représentées par l'ontologie UCC. Cela permet de mieux exploiter la date, d'extraire des informations plus importantes et de donner des résultats et des connaissances qui aident à créer de multiples applications, comme le montrent les illustrations suivantes. Nous donnons quelques exemples de ces questions :

Regroupement des utilisateurs par catégorie : Regroupement des utilisateurs isolés intéressés par la catégorie c à l'aide de la requête sqwrl:

```
Person(?p)^name(?p; ?n)^UserAccount(?u)^account(?u; ?p)^InterestT ree(?t)
^Interest(?i) ^ hasInterests(?u; ?t) ^ a setOf(?t; ?i)
^has startdate(?t; ?d1) ^ has enddate(?t; ?d2)
^has score(?i; ?s) ^ swrlb : greaterThan(?s; 0.15)
^swrlb : equal(?d1; 2015 - 11 - 11)
^swrlb : equal(?d2; 2015 - 12 - 11) ! query : select(?n; ?i; ?s)
```

Nous avons obtenu les résultats :

Ali s'intéresse à Science avec score = 0.18

Tina s'intéresse à Science avec score = 0.21

Lila s'intéresse à Science avec score = 0.41

...

Sofia s'intéresse aux News avec score = 0.30

Hakim s'intéresse aux News avec score = 0.22

Samir s'intéresse aux nouvelles avec score = 0.28

...

Kouki s'intéresse à la politique avec score = 0.24

...

Calculez le nombre d'utilisateurs intéressés par la catégorie news :

```
UserAccount(?u)^InterestT ree(?t)^Interest(politics)^has Interests(?u; news)
^a setOf(?t; politics)^has startdate(?t; ?d1)^has enddate(?t; ?d2)^has score(?i; ?s)
^swrlb : greaterThan(?s; 0.6) ^ swrlb : equal(?d1; 2015 - 11 - 11)
^swrlb : equal(?d2; 2015 - 12 - 11) ! query : count(?u)
```

19 utilisateurs sont intéressés par la categorie News du 11-11-2015 au 11-12-2015

Extraction des intérêts de chaque utilisateur: en utilisation de la requête:

```
Person(?p)^name(?p; ?n)^UserAccount(?u)^Post(?p)^has creator(?p; ?u)^Category(?c)
^topic(?p; ?c) ^ created(?p; ?time) ! query : select(?n; ?c; ?time)
```

Ahmed s'intéresse à Science au 28/01/2016 à 21h24.

Ahmed s'intéresse à Science au 28/01/2016 à 21h47.

...

Mona s'intéresse à l'art au 28/01/2016 à 21h52.

...

Omar s'intéresse à la politique au 29/01/2016 à 12h02.

...

En outre, en interrogeant l'ontologie UCC, il est possible d'obtenir des statistiques, tel que:

Requête 1:

```
Person(?p)^gender(?g; ?n)^UserAccount(?u)^account(?u; ?p)^InterestT ree(?t)^
Interest(sport)^has Interests(?u; ?t)^a setOf(?t; sport)^has startdate(?t; ?d1)^
has enddate(?t; ?d2)^has score(sport; ?s)^swrlb : greaterThan(?s; 0.20)^
swrlb : equal(?d1; 2015-11-11)^swrlb : equal(?d2; 2015-12-11) ! query : count(?p)
```

Requête 2:

Person(?p)^gender(?g; ?n)^swrlb : equal(?g; male)^UserAccount(?u)^account(?u; ?p)^
InterestT ree(?t)^Interest(sport)^has Interests(?u; ?t)^a setOf(?t; sport)^has startdate(?t; ?d_1)
^has enddate(?t; ?d_2)^has score(sport; ?s)^swrlb : greaterThan(?s; 0.20)
^swrlb : equal(?d_1; 2015-11-11)^swrlb : equal(?d_2; 2015-12-11) ! query : count(?p)

Après avoir appliqué la première requête, nous avons obtenu 48, et après avoir appliqué la seconde, nous avons eu 35, ce qui signifie que plus de 73% des utilisateurs s'intéressent au sport sont des hommes. De la même manière, nous pouvons obtenir d'autres résultats statistiques qui ont une relation avec l'âge, la profession, la zone géographique, etc.

Il est clair que le modèle que nous avons présenté dans cette thèse est beaucoup plus général que ses applications pratiques. Il est donc possible d'utiliser ce modèle en tenant compte de sa richesse structurelle et sémantique dans d'autres applications.

3.7. Comparaison entre le modèle UCC et les autres modèles

Dans cette section, nous comparons le modèle UCC avec plusieurs travaux de l’état de l’art. Cette comparaison est présentée dans le Tableau 3.9.

Caractéristiques	Modélisation du profil d'utilisateur			Catégorisation du contenu social			Les intérêts de l'utilisateur			Catégorisation des utilisateurs	
Travail	Caractéristique statique	Contenu	Caractéristique dynamique	Approche statique	Deep Learning	Catégorisation hiérarchique	Approche hiérarchique	Intérêts temporelle dynamique	Intérêts temporel et permanant	Approche basée sur les liens	Approche basée sur le contenu textuel
FOAF	*										
SIOC	*	*									
OUBO			*								
SWUM	*		*								
Lai et al. (2015)					*						
TUMS Tao et al., (2011)	*	*	*					*			
Piao and Breslin (2016)			*					*	*		
Jiang et al., (2015)			*				*	*	*		
Shah et al., (2018)			*					*			
Liu et al., (2018)			*	*		*					
Zhang et al., (2017)			*					*			*
Kontopoulos et al., (2014)				*							
Moussavi et al., (2016)										*	
UCC (notre modèle)	*	*	*	*	*	*	*	*	*		*

Tableau 3.9: comparaison entre le modèle UCC et les autres modelés de l'état de l'art

3.8. Conclusion

D’après les résultats obtenus dans ce chapitre, Il est clair que le modèle que nous avons présenté dans cette thèse est beaucoup plus général que ses applications pratiques. Il est donc possible d'utiliser ce modèle en tenant compte de sa richesse structurelle et sémantique dans d'autres applications telle que les systèmes de recommandation.

Conclusions générales et perspectives

Conclusions générales et perspectives

Le Web 2.0 joue un rôle de plus en plus important dans la vie sociale des personnes. Il contient de plus en plus d'informations au sujet de leur centre d'intérêts. Il a mis en place de nouveaux paradigmes, des outils et des services Web (comme les blogs, les wikis, les pratiques de marquage) qui sont largement acceptés dans les deux communautés publiques et scientifiques. Ces outils agissent généralement comme silos de données. L'exploitation de ces données est un véritable gain, pour plusieurs domaines.

En terminant notre thèse, il est relevant de rappeler nos hypothèses de départ tout en essayant de les confirmer. Dans les chapitres 2 et 3, nous avant adopter l'idée que le web sémantique fournit une structure plus facile et riche pour le web social. Il facilite les rapprochements et par conséquence améliore la pertinence des informations que les utilisateurs reçoivent du Web.

Plus précisément, nous avons proposé l'hypothèse suivante :

En s'appuyant sur la technologie du web sémantique et les techniques du prétraitement et de classification du texte, il est fortement probable de construire un modèle puissant qui permet la modélisation des intérêts des utilisateurs par la représentation sémantique des données non structurées issues des réseaux sociaux (le contenu et les données personnelles et sociales des internautes), la catégorisation des messages partager selon leurs thème, l'inférence et la modélisation des intérêts des utilisateurs et par conséquent, la catégorisation de ces utilisateurs, selon leurs centre d'intérêts.

Nous avons essayé de confirmer cette hypothèse dans des scenarios lies à l'innovation sur les réseaux sociaux.

Notre thèse a été conçue des trois chapitres suivants :

- ✓ Le premier chapitre a exposé tous les concepts et les travaux reliés à l'état de l'art en les répartissant en quatre parties :
  - Dans la partie 1, nous avons présenté l'aspect général des plateformes du Web social, et leurs caractéristiques en mettant l'accent sur les réseaux sociaux.
  - Dans la partie 2 du chapitre 1, nous avons montré comment les techniques du Web sémantique peuvent être utilisées conjointement pour gérer l'information affective associée à des données créées par les utilisateurs et les communautés en ligne. En particulier, les réseaux sociaux.

- Dans la partie 3 du même chapitre, nous avons expliqué comment utiliser les techniques du Data Mining, du Text Mining, de Traitements de langage naturel et de l'apprentissage automatique afin traiter les textes non formels issues des réseaux sociaux, et de fournir des modèles capables de classifier ces textes selon leurs intérêts, tout en basant sur la littérature existante.
- Dans la partie 4 du même chapitre, nous avons discuté les travaux reliés à la modélisation des intérêts des utilisateurs et les regroupant selon leurs intérêts.

✓ Dans le chapitre 2, nous avons introduit un nouveau modèle sémantique pour modéliser et analyser les données sociales ainsi que le texte en langage naturel des messages issus des réseaux sociaux, en exploitant les ontologies (SIOC, et FOAF) afin de coder les résultats dans un format courant sémantique.

✓ Dans le dernier chapitre, nous avons évalué le modèle proposé en modélisant un ensemble d'utilisateurs Facebook (leurs données personnelles, et social ainsi que le contenu partagé). Dans cette application nous avons catégorisé ces messages en utilisant deux classifications, une classification générale et une classification des news. Nous avons construit des arbres d'intérêts des utilisateurs, et regrouper ces derniers selon leurs intérêts. Nous avons obtenu de bons résultats en fournissant des connaissances et des informations utiles pour plusieurs application et domaines.

Champs d'application

Le modèle fourni en fin de cette thèse permet d'offrir une vue globale, spécifique et profonde sur les utilisateurs, et les communautés en ligne. Ceci ouvre une grande porte devant une variété de spécialistes comme les sociologues, les politiciens, les agents Marketing, etc. En l'utilisant, un maximum de connaissances et d'informations utiles soient disponible et exploitable.

Il est clair que le modèle que nous avons présenté dans cette thèse est bien plus général que ses applications concrètes. Il est donc possible d'utiliser ce modèle, en tenant compte de sa richesse structurelle et sémantique dans d'autres cas d'utilisation tels que :

- Les systèmes de diffusion de contenu pour mieux cibler un public plus large et plus précis ;
- La recherche d'experts à partir des sources fermées, i.e. au sein d'entreprises, pour mieux orienter la recherche en fonction d'une situation particulière ;
- L'élargissement des requêtes dans les systèmes de recherche

- Les systèmes de recommandation pour diversifier les recommandations et les rendre plus utiles en favorisant les centres d'intérêts, le comportement, etc.
- La détection des communautés en ligne,
- Les systèmes qui s'appuient sur la prédiction et la prise de décision.
- Les systèmes d'apprentissage automatiquement des intérêts de lecture des utilisateurs
- Les systèmes de filtrage automatique et de filtrage d'émail par contenu

Limitations

Le modèle proposé est basé sur des modèles d'apprentissage automatique dans lesquels sont choisis et construits par l'application utilisant le modèle proposé (UCC). Par conséquent, si l'application intègre des modèles d'apprentissage automatique peu performants, le modèle UCC donne de mauvais résultats, en particulier avec l'approche descendante proposée. À chaque niveau plus profond, la performance de la classification se dégrade encore plus.

De plus, nous limitons notre travail au texte où l'une des spécificités des réseaux sociaux est l'hétérogénéité des données (les publications peuvent être des images, des vidéos, du texte, etc.).

Perspectives

Dans nos travaux futurs, nous élargirons notre ensemble de données et nous nous concentrerons principalement sur la résolution de ces limitations :

- ✓ Les contenus multimédias issus des réseaux sociaux sont divers. Dans notre modèle proposé dans cette thèse, nous avons met la lumière que sur le traitement et la catégorisation des contenus textuels. Nous proposons aux futures travaux d'améliorer ce modèle en analysant les autres contenus multimédias par exemple : le prétraitement, le filtrage et la catégorisation des images, afin de mieux comprendre le comportement et le centre d'intérêts des utilisateurs et les communautés en ligne.
- ✓ Un autre point à évoquer, au lieu de catégoriser les messages textuels en utilisant la classification supervisée avec des catégories bien définies, les prochains travaux peuvent les catégorisés en générant des modelés non supervisé (Clustering) en se basant sur les caractéristiques de chaque groupe.
- ✓ Améliorer ce modèle par l'ajout d'autres aspects dynamiques, comme la modélisation des activités, la modélisation des besoins des utilisateurs et des communautés en ligne.
- ✓ L'amélioration des approches de prétraitements des textes brutes en se basant sur d'autres caractéristiques des textes issus des réseaux sociaux afin d'augmenter la performance des classifiers.

Références bibliographiques

Références

Abdelsadek, Y., Chelghoum, K., Herrmann, F., Kacem, I., & Otjacques, B. (2018). Community extraction and visualization in social networks applied to Twitter. *Information Sciences*, *424*, 204-223.

Abel, F. (2008). The benefit of additional semantics in folksonomy systems. In *Proceedings of the 2nd PhD workshop on Information and knowledge management* (pp. 49-56). ACM.

Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011). Semantic enrichment of twitter posts for user profile construction on the social web. In *Extended semantic web conference* (pp. 375-389). Springer, Berlin, Heidelberg.

Akaichi, J., Dhouioui, Z., & Pérez, M. J. L. H. (2013, October). Text mining facebook status updates for sentiment classification. In *2013 17th International conference on system theory, control and computing (ICSTCC)* (pp. 640-645). IEEE.

Al-Ayyoub, M., Essa, S. B., & Alsmadi, I. (2015). Lexicon-based sentiment analysis of Arabic tweets. *IJSNM*, *2*(2), 101-114.

Al-Smadi, M., Talafha, B., Al-Ayyoub, M., & Jararweh, Y. (2018). Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. *International Journal of Machine Learning and Cybernetics*, 1-13.

AlSuwaidan, L., & Ykhlef, M. (2018). Interest-based clustering approach for social networks. *Arabian Journal for Science and Engineering*, *43*(2), 935-947.

Amiri, H., & Chua, T. S. (2012, February). Mining slang and urban opinion words and phrases from cQA services: an optimization approach. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 193-202). ACM.

Bachelet, R. (2014). Réseaux sociaux, Centrale Lille, 40p

Benkhelifa, R., & Laallam, F. Z. (2016). Facebook Posts Text Classification to Improve Information Filtering. In *WEBIST (1)*(pp. 202-207).

Benkhelifa, R., & Laallam, F. Z. (2018, February). Opinion extraction and classification of real-time youtube cooking recipes comments. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 395-404). Springer, Cham.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, *284*(5), 28-37.

Bilski, A. (2011). A review of artificial intelligence algorithms in document classification. *International Journal of Electronics and Telecommunications*, *57*(3), 263-270.

Breslin, J. G., Harth, A., Bojars, U., & Decker, S. (2005). Towards semantically-interlinked online

communities. In *European semantic web conference* (pp. 500-514). Springer, Berlin, Heidelberg.

Budak, C., Kannan, A., Agrawal, R., and Pedersen, J. (2014) 'Inferring user interests from microblogs'. Technical Report, MSR-TR-201468.

Camarinha-Matos, L. M., & Afsarmanesh, H. (2004). A multi-agent based infrastructure to support virtual communities in elderly care. *International Journal of Networking and Virtual Organisations*, *2*(3), 246-266.

Contratres, F. G., Alves-Souza, S. N., Filgueiras, L. V. L., & DeSouza, L. S. (2018, March). Sentiment analysis of social network data for cold-start relief in recommender systems. In *World Conference on Information Systems and Technologies*(pp. 122-132). Springer, Cham.

DCTerms | www.dublincore.org/documents/dcmi-terms

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization.

Echarte, F., Astrain, J. J., Córdoba, A., & Villadangos, J. E. (2007). Ontology of Folksonomy: A New Modelling Method. *SAAKM*, *289*, 36.

Faqeeh, M., Abdulla, N., Al-Ayyoub, M., Jararweh, Y., & Quwaider, M. (2014, August). Cross-lingual short-text document classification for facebook comments. In *2014 International Conference on Future Internet of Things and Cloud* (pp. 573-578). IEEE.

Faralli, S., Stilo, G., & Velardi, P. (2017). Automatic acquisition of a taxonomy of microblogs users' interests. *Journal of Web Semantics*, *45*, 23-40.

Feng, G., Li, S., Sun, T., & Zhang, B. (2018). A probabilistic model derived term weighting scheme for text classification. *Pattern Recognition Letters*, *110*, 23-29.

Feng, S., Wang, Y., Liu, L., Wang, D., & Yu, G. (2019). Attention based hierarchical LSTM network for context-aware microblog sentiment classification. *World Wide Web*, *22*(1), 59-81.

FOAF | www.foaf-project.org

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI magazine*, *13*(3), 57-57.

Gruber, T. (2005). Folksonomy of ontology: A mash-up of apples and oranges.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.

hcard Microformat | www.microformats.org/wiki/hcard

Ho, T., & Do, P. (2015, October). Discovering Communities of Users on Social Networks Based on Topic Model Combined with Kohonen Network. In *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)* (pp. 268-273). IEEE.

Hong, L., Doumith, A. S., & Davison, B. D. (2013, February). Co-factorization machines: modeling user interests and predicting individual decisions in twitter. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 557-566). ACM.

Huang, M., Zou, G., Zhang, B., Liu, Y., Gu, Y., & Jiang, K. (2018). Overlapping community detection in heterogeneous social networks via the user model. *Information Sciences*, *432*, 164-184.

Jaffali, S., Jamoussi, S., & Hamadou, A. B. (2014, September). Grouping like-minded users based on text and sentiment analysis. In *International Conference on Computational Collective Intelligence* (pp. 83-93). Springer, Cham.

Jaffali, S., Jamoussi, S., Hamadou, A. B., and Smaili, K. (2014) 'Clustering and Classification of Like-Minded People from Their Tweets'. In Data Mining Workshop (ICDMW), 2014 IEEE International Conference, pp. 921-927. IEEE.

Jarwar, M. A., Abbasi, R. A., Mushtaq, M., Maqbool, O., Aljohani, N. R., Daud, A., ... & Chong, I. (2017). CommuniMents: A framework for detecting community based sentiments for events. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *13*(2), 87-108.

Jiang, B., & Sha, Y. (2015). Modeling temporal dynamics of user interests in online social networks. *Procedia Computer Science*, *51*, 503-512.

Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137-142). Springer, Berlin, Heidelberg.

Kang, J., and Lee, H. (2017). 'Modeling user interest in social media using news media and wikipedia'. Information Systems, No. 65,pp. 52-64.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), 59-68.

Kermanidis, K. L., & Maragoudakis, M. (2013). Political sentiment analysis of tweets before and after the Greek elections of May 2012. *International Journal of Social Network Mining*, *1*(3-4), 298-317.

Khemmarat, S., Saha, S., Song, H. H., Baldi, M., and Gao, L. (2014) 'On understanding user interests through heterogeneous data sources'. In International Conference on Passive and Active Network Measurement pp. 272-274. Springer International Publishing.

Kim, H. L., Breslin, J. G., Yang, S. K., & Kim, H. G. (2008, March). Social semantic cloud of tag: Semantic model for social tagging. In *KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications* (pp. 83-92). Springer, Berlin, Heidelberg.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

Kundi, F. M., Ahmad, S., Khan, A., & Asghar, M. Z. (2014). Detection and scoring of internet slangs for sentiment analysis using SentiWordNet. *Life Science Journal*, *11*(9), 66-72.

Laloë, T. (2009). *Sur quelques problèmes d'apprentissage supervisé et non supervisé* (Doctoral dissertation, Université Montpellier II-Sciences et Techniques du Languedoc).

Lan, M., Sung, S. Y., Low, H. B., & Tan, C. L. (2005, July). A comparative study on term weighting schemes for text categorization. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* (Vol. 1, pp. 546-551). IEEE.

Li, X., Pang, J., Mo, B., Rao, Y., & Wang, F. L. (2016, April). Deep neural network for short-text sentiment classification. In *International Conference on Database Systems for Advanced Applications* (pp. 168-175). Springer, Cham.

Lin, C., Li, Y., Chang, T. W., and Kinshuk. (2017). 'Analysing user retweeting behaviour on microblogs: prediction model and inuencing features'. International Journal of Computational Science and Engineering, Vol. 15, No. 3, pp. 176-187.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Liu, Y., Wang, J., Jiang, Y., Sun, J., and Shang, J. (2018). 'Identifying impact of intrinsic factors on topic preferences in online social media: A nonparametric hierarchical Bayesian approach'. Information Sciences, No. 423, pp. 219-234.

Lundborg, A. (2017). Text classification of short messages. *LU-CS-EX 2017-14*.

Madan, M., Dave, M., & Chopra, M. (2018). Analyzing Online Groups or the Communities in Social Media Networks by Algorithmic Approach. In *ICT Based Innovations* (pp. 113-126). Springer, Singapore.

Manrique, R., Herazo, O., and Mario, O. (2017). 'Exploring the Use of Linked Open Data for User Research Interest Modeling'. In Colombian Conference on Computing, pp. 3-16. Springer, Cham.

Maret, P., Hammond, M., & Calmet, J. (2004, October). Virtual knowledge communities for corporate knowledge issues. In *International Workshop on Engineering Societies in the Agents World* (pp. 33-44). Springer, Berlin, Heidelberg.

Meaning of a Tag (MOAT) | moat-project.org

Mendes, P. N., Passant, A., Kapanipathi, P., & Sheth, A. P. (2010, August). Linked open social signals. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 224-231). IEEE.

MESH user ontology | www.mesh-ip.eu/upload/MESH_user_ontology.zip

Microformats | www.microformats.org

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web semantics: science, services and agents on the World Wide Web*, *5*(1), 5-15.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. ICLR Workshop, 2013.

Miles, A., & Bechhofer, S. (2009). SKOS simple knowledge organization system reference. *W3C recommendation*, *18*, W3C.

Moosavi, S. A., Jalali, M., Misaghian, N., Shamshirband, S., and Anisi, M. H. (2016) 'Community detection in social networks using user frequent pattern mining'. Knowledge and Information Systems, pp. 1-28.

Moualek Djaloul Youcef. *Deep Learning pour la classification des images*. 2017. Thèse de doctorat. 07-03-2017.

Naaman, M., Boase, J., & Lai, C. H. (2010, February). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 189-192). ACM.

Newman, R., Ayers, D., and Russell, S., (2005). Tag Ontology, http://www.holygoat.co.uk/owl/redwood/0.1/tags/.

O'Connor, M. J., & Das, A. K. (2009, October). SQWRL: A Query Language for OWL. In *OWLED* (Vol. 529).

Omitola, T., Ríos, S. A., & Breslin, J. G. (2015). *Social semantic web mining*. Morgan & Claypool Publishers.

Online Presence Ontology (OPO) | www.online-presence.net/ontology.php

Orlandi, F., Breslin, J., and Passant, A. (2012) 'Aggregated, interoperable and multi-domain user profiles for the social web'.In Proceedings of the 8th International Conference on Semantic Systems, pp. 41-48. ACM.

Pasolini, R. (2015). *Learning Methods and Algorithms for Semantic Text Classification across*

Multiple Domains (Doctoral dissertation, alma).

Passant, A. (2008). LODr - A Linking Open Data Tagging System. In Proceedings of the First Social Data on the Web Workshop (SDoW2008), Karlsruhe, Germany.

Passant, A., & Laublet, P. (2008). Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. *LDOW*, *369*.

Piao, G., & Breslin, J. G. (2018). Inferring user interests in microblogging social networks: a survey. *User Modeling and User-Adapted Interaction*, *28*(3), 277-329.

Piao, G., and Breslin, J. G. (2016) 'Exploring dynamics and semantics of user interests for user modeling on Twitter for link recommendations'. In Proceedings of the 12th International Conference on Semantic Systems. pp. 81-88. ACM.

Relationship Vocabulary | www.vocab.org/relationship

Rheingold, H. (2000). The virtual community: Homesteading on the electronic frontier. Cambridge, MASS: The MIT Press.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851-860). ACM.

Schema.org | www.schema.org

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Sebastiani, F. (2005). Text categorization text mining and its applications. *Text Mining and its Applications, A*, 109-129.

Shah, B., Verma, A. P., & Tiwari, S. (2018). User Interest Modeling from Social Media Network Graph, Enriched with Semantic Web. In *Proceedings of International Conference on Computational Intelligence and Data Engineering* (pp. 55-64). Springer, Singapore.

Stan, J. (2011). *A semantic framework for social search* (Doctoral dissertation, Saint-Etienne).

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1555-1565).

Tao, K., Abel, F., Gao, Q., and Houben, G. J. (2011) 'Tums: twitter-based user modeling service'.In Extended Semantic Web Conference, pp. 269-283. Springer Berlin Heidelberg.

TextRazor | www.textrazor.com

Uttarwar, M., & Bhute, Y. (2013). A Review on Customizable Content-Based Message Filtering from OSN User Wall.

Vanetti, M., Binaghi, E., Ferrari, E., Carminati, B., & Carullo, M. (2013). A system to filter unwanted messages from osn user walls. *IEEE Transactions on Knowledge and data Engineering*, *25*(2), 285-297.

Véronis, J., & Guimier de Neef, E. (2006). Le traitement des nouvelles formes de communication écrite. *Compréhension automatique des langues et interaction*, 227-248.

Wellman, B. (1996, April). For a social network analysis of computer networks: a sociological perspective on collaborative work and virtual community. In *Proceedings of the 1996 ACM SIGCPR/SIGMIS conference on Computer personnel research* (pp. 1-11). ACM.

Wellman, B. (2001). Computer networks as social networks. *Science*, *293*(5537), 2031-2034.

Weston, J., & Watkins, C. (1998). *Multi-class support vector machines* (pp. 98-04). Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May.

XFN microformat | www.gmpg.org/xfn

Xu, Z., Ru, L., Xiang, L., and Yang, Q. (2011) 'Discovering user interest on twitter with a modified author-topic model'. In Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Vol. 01, pp. 422-429. IEEE Computer Society.

Yamak, Z., Saunier, J., & Vercouter, L. (2018). SocksCatch: Automatic detection and grouping of sockpuppets in social media. *Knowledge-Based Systems*, *149*, 124-142.

Yeh, I., Karp, P. D., Noy, N. F., & Altman, R. B. (2003). Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics*, *19*(2), 241-248.

Zarrinkalam, F., Fani, H., Bagheri, E., & Kahani, M. (2016, March). Inferring implicit topical interests on twitter. In *European Conference on Information Retrieval* (pp. 479-491). Springer, Cham.

Zarrinkalam, F., Fani, H., Bagheri, E., Kahani, M., and Du, W. (2015) 'Semantics-enabled user interest detection from twitter'. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference, Vol. 1, pp. 469-476. IEEE.

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1-4), 43-52.

Zhang, Y., Wang, X., Xiao, C., Sun, Y., & Qiao, Y. (2017). Temporal dynamic recommendation based on data imputation through association analysis. *International Journal of Computational Science and Engineering*, *15*(1-2), 130-137.

Zhang, Z., He, Q., Gao, J., & Ni, M. (2018). A deep learning approach for detecting traffic

accidents from social media data. *Transportation research part C: emerging technologies*, *86*, 580-596.

Zhao, A. P., Yu, L., and Yang, W. M. (2016). 'Semantically structured service community discovery: based on relationship and functionality'. International Journal of Computational Science and Engineering, Vol. 13, No. 3, pp. 233-245.

Zhao, J., Dong, L., Wu, J., & Xu, K. (2012, August). Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1528-1531). ACM.