UNIVERSITE KASDI MERBAH OUARGLA

Faculté des Sciences, de la Technologie et des Sciences de la Matière

Département des Mathématiques et d'Informatique

Mémoire

MASTER ACADEMIQUE

Domaine : Mathématiques et Informatique.

Filière : Informatique académique.

Spécialité : Informatique Fondamentale.

Présenté par : **M : Mohammed Tayeb KERROUCHE**

M : Ali BETTAYEB

.

Thème

Application pour La protection des mineurs contre les dangers des réseaux sociaux basé sur les techniques de bio-inspirés

Soutenu publiquement

le :../6/2016.

Proposé par : **Mr . BOUHANI Abdelkader**

Devant le jury

Année Universitaire : 2015 /2016

Dédicaces

Du profond de mon cœur, je dédie ce travail

A l'esprit de mon père Que Dieu bénisse son âme ,

A ma mère, et ma femme, pour votre soutien moral

A mes enfants Hazem et Maya,

A mes frères, sœurs,

A TOUS MES AMIES…

Tayeb

Dédicaces

Du profond de mon cœur, je dédie ce travail

A l'esprit de mon père Que Dieu bénisse son âme ,

A ma mère, et ma femme, pour votre soutien moral

A mes frères, sœurs,

A TOUS MES AMIES…

Tayeb et ali

Remerciements

Je remercie tout d'abord notre Dieu qui m'a donné la force et la volonté pour élaborer ce travail.

J'adresse mes vifs remerciements à mon encadreur

Mr. BOUHANI Abdelkader

qui m'a aidé durant mon travail et par sa patience et ses précieux conseils dont Il m'a entouré.

Sans son aide, mon travail n'aurait pas vu la lumière.

Je remercie vivement les membres du jury qui m'ont fait l'honneur d'accepter de juger notre travail.

ma reconnaissance aussi à tous ceux qui ont collaboré à ma formation en particulier les enseignants du département d'Informatique, de l'université KASDI MARBAH OUARGLA surtout Dr. KHERFI Mohammed Lamine.,mahjoub Mohammed bachir

Je remercie également tous ceux qui ont participé de près ou de loin à élaborer ce travail.

Tayeb et Ali

Table des matières

Bibliographie

Liste des figures

I. Introduction Générale :

Un monde sans Réseaux Sociaux est aujourd'hui presque inimaginable et ce, particulièrement pour les enfants et adolescents qui ont recours à cet outil de plus en plus fréquemment, que ce soit pour consulter des profils , envoyer un message, « chatter » avec des amis, télécharger de la musique ou encore se documenter pour un travail scolaire. Malheureusement, les Réseaux Sociaux ne comportent pas que des avantages. C'est également un lieu où abus et excès en tout genre se rencontrent. Face à ces dérives, il est très rapidement apparu indispensable de protéger les mineurs, contre ces dangers.

D'autre coté en informatique, La bio-inspiration est un changement de paradigme qui amène les ingénieurs à s'inspirer de la nature pour développer de nouveaux systèmes artificiels. Ainsi plusieurs solutions informatique étaient inspirée de la nature en mémitant des vivants, comme les oiseaux, les poissons, les termites, les colonies de fourmis, les insectes ; La problématique est de construire une application permettant d'assurer la sécurité des mineurs contre les dangers des réseaux sociaux, en utilisant une méthode inspirée de la nature.

Nous proposons dans ce qui suit une application de protection des utilisateurs mineurs de Facebook basé sur une approche bio-inspiré l'optimisation par essaims particulaires binaire "BPSO" en anglais, hybridé par la technique K plus proche voisins "KNN" en anglais. Nous commençons dans le 1ére chapitre par présenter le Facebook, leurs spécificités et les différents aspects sur lesquels ils sont basés, et ainsi nous présentons la problématique et les principaux dangers auxquels sont confrontés actuellement les internautes mineurs. Le 2éme chapitre présente les techniques employées dans les différentes phases du processus Catégorisation Automatique de Textes :la "PSO", son origine et idée de base, était le fruit des études faites par James Kennedy et Russ Eberhart sur le comportement de Floking en 1995, ensuite on a étudié sa description formelle, ses avantages et ses inconvenants, et ses variantes, en fin on exposel'algorithme de BPSO hybridé parKNN ainsi que les différentes techniques utilisées. Dans le3éme chapitrenous présentons la conception de notre application et une présentation de langage UML selon notre besoins, et dans le 4éme chapitre on introduit l'implémentation del'application ainsi que les résultats obtenus lors de la validation.

Chapitre 1 :

Facebook et les réseaux sociaux

I. Chapitre 1 :Facebook et les réseaux sociaux

I.1. Introduction :

À la fin des années 1990, les réseaux sociaux sont apparus sur Internet réunissant des personnes via des services d'échanges personnalisés. Ainsi, un réseau social peut être considéré comme un ensemble de personnes réunies par un lien social. L'office québécois de la langue française (Office québécois de la langue française 2012) définit le réseau social comme étant une « communauté d'internautes reliés entre eux par des liens, amicaux ou professionnels, regroupés ou non par secteurs d'activités, qui favorisent l'interaction sociale, la création et le partage d'informations ».

Le premier réseau social est lancé en 1997. C'était *SixDergrees.com*. Il permet déjà à l'utilisateur de créer un profil et lister ses amis. Ce n'est pas la première fois qu'on utilise le concept de liste d'amis, cela existe depuis longtemps dans les sites de rencontres tels que *icq.com* ou *aim.com* où on peut ajouter des amis avec leurs noms, mais ce qui a changé c'est que la liste d'amis d'un utilisateur est devenue visible à tous ses amis aussi.

Les réseaux sociaux permettent aux utilisateurs de rester en contact avec des amis, de discuter et de partager avec eux des informations. Ils leurs fournissent des outils d'interaction et de communication qui leurs permettent d'agrandir leurs réseaux d'amis et d'interagir via des applications tierces.

I.2. Évolution des réseaux sociaux :

Le concept de réseau social professionnel fait réellement surface à partir de 2001, avec *Ryse.com*.À partir de 2003, nous assistons à la naissance de beaucoup de réseaux sociaux se dirigeant vers le même concept élaboré par *Friendster* tout en ciblant des communautés bien spécifiques selon leurs emplacements démographiques, religions ou intérêts. L'introduction de *Facebook* en 2004 marque le début d'une période de réseautage importante. Depuis, nous assistons sans cesse à la naissance de nouveaux réseaux innovateurs, dont les plus récents sont*Twitter,YouTube,LinkedIn, Instagram,Myspace,Google+*.

I.3. Pourquoi utilise-t-on les réseaux sociaux :

Parmi les raisons de l'utilisation des réseaux sociaux. On peut extraire :

- Créer de nouvelles relations et augmenter la taille de son réseau.
- Retrouver des camarades de classe et d'anciens amis.
- Partager ces passions avec d'autres personnes qui ont les mêmes passions.
- Organiser des événements.
- Faire du réseautage pour un but professionnel (trouver un emploi).

I.4. Le Facebook :

Développé en 2004 par l'étudiant de l'Université de Harvard Mark Zuckerberg, Facebook est un site de réseau social utilisé par plus de 800 millions d'utilisateurs actifs dans tous les pays de la planète, à ce jour dans 70 langues. Âge minimum du site est 13, mais les adolescents ne représentent qu'une population minoritaire sur Facebook. Il est utilisé par un beaucoup d'adultes, y compris certainement les parents. Mais pas seulement des individus - de Facebook également utilisé par les entreprises, les organisations et les gouvernements partout dans le monde, envoyer des messages marketing, rechercher des fonds de bienfaisance et de communiquer avec les clients et les électeurs .

Facebook est certainement pas le seul site de réseautage social. Il y a des milliers d'eux, basés partout dans le monde, certains sites sociaux d'intérêt général pour les personnes dans un pays spécifique et certains des groupes d'intérêts spécifiques dans de nombreuses catégories -les étudiants, les amateurs de sport, les amateurs de films, des cuisiniers, des voyageurs, gammes, amateurs de musique, etc. Certains sites sociaux sont conçus pour être utilisés sur des ordinateurs, certain juste pour mobile Téléphones. Facebook est accessible à la fois. [https://fr.wikipedia.org/wiki/Facebook].

I.5. Principaux Termes de Facebook :

Facebook possède un glossaire de termes lui étant propre. Nous exposons dans cette partie certains de ces termes jugés importants dans la compréhension de son fonctionnement.

- **Amis (*Friends*)**: Ils représentent les personnes avec qui un utilisateur peut entrer en contact et partager des informations (images, textes, vidéos, etc.). Tout le

principe de *Facebook* repose sur cette notion de partage d'informations entre les réseaux d'amis. Une fois la personne recherchée trouvée, l'utilisateur doit cliquer sur le bouton « Ajouter à mes amis ». À ce moment, une demande d'ajout à la liste d'amis sera envoyée à cette personne. Une fois la confirmation obtenue, cette personne devient alors ami(e) avec l'utilisateur et apparaît dans sa liste d'amis *Facebook*. Il est à noter que les paramètres de confidentialité peuvent limiter la recherche de certains utilisateurs. À tout moment un utilisateur peut retirer un ami de sa liste d'amis. Il peut aussi décider de « bloquer » un ami.

- **Paramètres du compte *(Account settings)***: Les paramètres du compte d'un utilisateur permettent de gérer les préférences de base pour son compte. Il peut entre autres modifier son nom, son adresse électronique de connexion, son mot de passe, ses préférences de notification ou ses fonctions de sécurité supplémentaires.
- **Paramètres de confidentialité (*Privacy settings*)** : Les paramètres de confidentialité permettent à l'utilisateur de gérer les options de confidentialité de son compte Facebook. Il peut, par exemple, indiquer qui pourra lui envoyer des demandes d'ajout d'amis et des messages. Pour toute autre information qu'il partage sur Facebook, il a la possibilité de choisir les personnes qui recevront chacune de ses publications. Normalement, l'utilisateur a le choix entre quatre options :Public, Amis, Moi uniquement, Personnaliser (listes des personnes que l'utilisateur a choisi d'inclure ou d'exclure).
- **Journal (*Timeline*)**: Le journal d'un utilisateur (anciennement appelé profil) est le recueil qui englobe l'ensemble des photos, interactions, publications, expériences et activités de l'utilisateur. Il affiche les évènements par ordre chronologique. On y retrouve sa photo, son nom et prénom, sa date de naissance, sa ville de naissance, sa ville actuelle, son niveau d'éducation, le nom de son employeur actuel, son orientation politique, sa religion, ses intérêts musicaux, ses citations favorites et bien plus. L'utilisateur peut choisir de n'afficher qu'une partie de ces renseignements, et ce, grâce aux paramètres de sécurité du compte. Il peut également presque tout cacher ou tout afficher.
- **Mur (*Wall*)** : Le mur est un espace du journal où l'utilisateur peut publier et échanger du contenu avec ses amis. Cet échange peut être sous forme de messages textes, d'images, de vidéos ou de liens vers du contenu sur Internet. Toute activité exécutée par l'utilisateur et ses amis sera annoncée sur son mur. Cependant,

Facebook donne la possibilité à l'utilisateur de restreindre l'accès à son mur et de limiter l'accès à seulement quelques amis. En tout temps, l'utilisateur peut choisir de masquer une actualité visible sur son mur, et ce, en appuyant sur le bouton « X » apparaissant à côté de l'actualité. Cette dernière ne sera alors plus visible.

- **Messages (*Messages*)**: Les messages jouent un rôle majeur dans l'échange des messages privés, des discussions instantanées, des messages électroniques et des textes avec des amis. Les utilisateurs de*Facebook* peuvent par exemple échanger des messages à travers la messagerie instantanée (Chat). Pour lancer une discussion, il suffit de cliquer sur le nom de l'ami avec qui un usager souhaite clavarder pour que *Facebook* ouvre une fenêtre de discussion. *Facebook*met également à la disposition de ses « échangeurs de messages » un ensemble d'icônes leur permettant de révéler leur humeur lors d'une discussion.
- **Statut (*Status*)** : C'est une fonction qui permet à l'utilisateur de faire un commentaire ou d'exprimer un avis. Semblable à un *tweet*, un statut est généralement court. Il exprime un point de vue sans entrer trop dans les détails.
- **Photos (*Photos*)** : C'est une fonction qui permet de partager des images et de marquer (*tag*) les personnes qui y figurent. Un utilisateur *Facebook* peut insérer une photo de profil, télécharger des photos, publier une photo sur un mur, ajouter des photos dans les messages et les groupes, créer et gérer un album de photos, etc. L'album de photos est défini par un titre et une description. Son créateur peut préciser qui peut le voir. Chaque album comporte une limite maximale de 1000 photos. Il n'existe cependant pas de limite au nombre d'albums autorisés. Toute photo de l'album peut être « aimée », « commentée » et « partagée » par les amis.
- **Appel vidéo (*Vidéo Chat*)** : L'appel vidéo est une fonctionnalité de *Facebook* qui permet de communiquer avec un ami avec l'image et le son. C'est une option disponible dans les fenêtres de clavardage du réseau social. Si l'ami ne répond pas à l'appel vidéo, il est alors possible de lui laisser un message vidéo enregistré. La date et l'heure de l'appel seront enregistrées dans l'historique, mais contrairement à la discussion instantanée, l'appel vidéo ne sera pas sauvegardé par le réseau. Seuls les amis de l'utilisateur peuvent l'appeler par vidéo.
- **Groupe (*Group*)**: Les utilisateurs de *Facebook* peuvent se regrouper autour de sujets d'intérêts communs sous formes de *groupes* (Ex : anciens étudiant d'un

lycée). Tout utilisateur peut créer un groupe sur *Facebook*. Il existe trois options de confidentialité :

- Ouvert : tout utilisateur de *Facebook* peut voir le groupe et le rejoindre.
- Fermé : tous les utilisateurs de *Facebook* peuvent voir le nom et les membres d'un groupe ainsi que les personnes invitées à rejoindre ce groupe, mais seuls les membres peuvent accéder aux publications correspondantes.
- Secret : ces groupes n'apparaissent pas dans les résultats de recherche et les personnes qui n'en sont pas membres ne peuvent rien voir, ni le contenu, ni la liste des membres. Le nom du groupe ne s'affichera pas sur le profil (journal) de ses différents membres. Pour rejoindre un groupe secret, il faut être ajouté par un membre existant de ce groupe.

- **Fil d'actualité (News feed)** : C'est la liste permanente des mises à jour sur la page d'accueil de l'utilisateur. Il affiche les publications des amis et les actualités des pages qu'il suit.
- **Page (*Page*)**: Les pages permettent aux entreprises, marques et personnalités de communiquer avec des personnes sur *Facebook*. Les administrateurs de ces pages peuvent publier des informations et mettre à jour leur fil d'actualité à l'attention des personnes qui les « suivent ». Les pages sont destinées à un usage professionnel et officiel : elles permettent aux organisations, entreprises, célébrités ou groupes de musique d'être présents sur*Facebook*. Elles permettent à leurs propriétaires de partager leurs histoires et de communiquer avec les usagers. Le but est d'intéresser et d'élargir le public grâce à des publications régulières. Par conséquent, les personnes qui aiment ces pages recevront les mises à jour dans leur fil d'actualité. Seul le représentant officiel d'un organisme, d'une entreprise, d'un artiste ou d'un groupe musical est autorisé à créer une page. Les pages ne sont pas des comptes *Facebook* distincts, elles partagent les mêmes informations de connexion qu'un profil personnel. Une fois qu'un administrateur a configuré une page au sein de son profil, il peut ajouter des administrateurs supplémentaires pour l'aider à la gérer.
- **Article (*Note*)** : Un article permet à un utilisateur de s'exprimer dans un format enrichi. Il peut contenir du texte, des images, des vidéos et des liens vers d'autres pages Web. Les articles sont normalement publics. Tout usager peut les consulter et les commenter.

- **J'aime (*Like*)**: Le bouton « J'aime » de *Facebook* permet aux internautes de donner un avis positif sur le contenu d'une publication et de s'associer à des sites Web favoris.
- **Évènement (*Event*) :***Facebook*permet à l'utilisateur d'organiser des évènements et d'y inviter ses amis. Pour cela, il doit déterminer la date, l'heure, l'emplacement et la nature de l'évènement.
- **Notification (*Notification*)**: C'est un message électronique à propos des mises à jour de l'activité sur *Facebook.*

I.6. Utilise les enfants au Facebook :

Pour autant de raisons que les adultes. La recherche de psychologues et de sociologues nous montre qu'ils utilisent les sites de réseautage social pour :

- Socialiser ou «traîner» avec leurs amis, pour la plupart des amis de pièces à école.
- Journée-à-jour des nouvelles de leurs amis, connaissances, parents, et les groupes de pairs.
- Collaboration sur le travail scolaire.
- Validation ou de soutien affectif.
- L'expression de soi et l'exploration de l'identité et de la formation qui se produit dans développement de l'adolescent.
- Qu'est-ce que les sociologues appellent «l'apprentissage informel», ou à l'extérieur d'apprentissage des paramètres formels comme l'école, y compris l'apprentissage des normes sociales et l'alphabétisation sociale.
- L'apprentissage des compétences techniques de l'ère numérique, que beaucoup de gens d'affaires se sentent sont essentiels au développement professionnel.
- Découvrir et explorer les intérêts, à la fois académique et professionnelle future intérêts.
- L'apprentissage sur le monde au-delà de leur maison immédiate et l'école environnements.
- L'engagement civique - participé à des causes qui ont un sens pour eux.[Web, 01].

I.7. Sécurité sur Facebook :

Tout comme les communautés dans le monde physique, pas de site de réseautage social, virtuelmonde, jeu en ligne, ou tout autre service de médias sociaux peut fournir une garantie de100% de sécurité, Facebook inclus. Pourquoi ? Parce que cela est le Web social, etla sécurité dépend beaucoup sur le comportement des utilisateurs vers l'autre. Facebook fournit des fonctions et de l'éducation pour ses utilisateurs sécurité et de confidentialité. Les parentsbénéficiés de la visite Center de Facebook, une ressource complète pour tousLes utilisateurs de Facebook avec des zones spéciales pour les adolescents, les parents, les éducateurs, et le droitmise en vigueur. Cette information en place la sécurité et ce guide sont importants pourla raison même que le «produit» de Facebook est produit par ses utilisateurs. Les parents ont besoinà savoir que, sur le Web social, la sécurité est une responsabilité partagée - une négociation Donc, la réponse courte à cette question est que, dans ce nouveau média, très sociall'environnement, la sécurité d'un utilisateur dépend de l'utilisateur autant sur le site. C'est pourquoi les parents doivent être informés et de garder les lignes de communication avec leur enfants grands ouverts - parce que les jeunes, comme tous les utilisateurs de Facebook, sont constammentcommunication, l'affichage et le partage de contenu sur le site.[Web, 02]

I.8. Principaux dangers dus aux réseaux sociaux :

Les médias sociaux présentent aussi des risques :

- prise de connaissance insuffisante de l'accessibilité des commentaires, photos, etc. et des risques d'utilisation frauduleuse des données qui en découlent. Les images qui circulent sur le Net sont ineffaçables ;
- risque d'inattention pour les jeunes qui font leurs devoirs sur ordinateur tout en étant connectés à un réseau social ;
- contacts indésirables et agressions sexuelles : les pédophiles peuvent utiliser les réseaux sociaux pour entrer en contact avec des victimes potentielles ;
- risque d'être ridiculisé, insulté ou harcelé par d'autres utilisateurs.

I.9. CONCLUSION :

Maintenant, il devrait être clair que Facebook est un site géant des réseaux sociaux fournissant une grande gamme diversifiée de services et de fonctionnalités. Il est aussi le reflet de et plate-forme pour les pensées, les actions, la créativité et l'apprentissage d'une grande section de l'humanité. Comment les gens utilisent le site est très individuel, et en gardant leur expériences sur le site positif dépend beaucoup de la façon dont ils l'utilisent et interagir avec les autres sur elle. Cela est tout aussi vrai pour les jeunes utilisateurs de Facebook comme pourles grandes personnes. Parce que l'utilisation de Facebook est basé sur les noms et les identités réelles, il est directement lié à «Vraie vie» - dans le cas des jeunes, la plupart du temps la vie et les relations école. Alorscomme dans la vie en ligne, les enfants ont besoin de l'aide de leurs parents alors qu'ils naviguent à la fois l'adolescence et le Web social. Vous pouvez les aider à comprendre.

Chapitre 2 :

Codage des textes

Etat de l'art

ii. Codage des textes : Etat de l'art

ii.1. Introduction:

La transformation ou le codage de ces documents est une préparation à « l'informatisation »de ces derniers, chaque type de documents comme les images, les vidéos et notamment lestextes dispose de ses propres techniques de codage.

Plusieurs approches de représentation des documents textuels ont été proposées dans ce contexte, la plupart étant des méthodes vectorielles.
Les principales méthodes de représentation de textes n'utilisent pas d'informationgrammaticale ni d'analyse syntaxique des mots : seule la présence ou l'absence de certains
mots est porteuse d'informations.

Un état de l'art des différentes approches de représentation et codage de textes est développédans ce chapitre.

Un texte est plutôt considéré comme une séquence de mots (unmot lui même étant une séquence de caractères) et représenté dans un espace de mots dont ladimension est plus grande que celle du caractère (Le nombre de caractères possibles est limitémais en revanche le nombre de mots qu'on peut avoir est énorme), mais dont chaquedimension est beaucoup plus informative.

ii.2. La catégorisation de textes :

La Catégorisation de Textes (C.T) est le processus qui consiste à assigner une ou plusieurscatégories parmi une liste prédéfinie à un document. L'objectif du processus est d'êtrecapable d'effectuer automatiquement les classes d'un ensemble de nouveaux textes. Principalement, les algorithmes de catégorisation s'appuient sur des méthodes d'apprentissagequi, à partir d'un corpus d'apprentissage, permettent de catégoriser de nouveaux textes. Cetype de méthodes sont dites inductives car elles induisent de la connaissance à partir desdonnées en entrée (les textes) et des sorties (leurs classes).

Pour réaliser l'opération de catégorisation automatique de textes comme nous l'avons défini, la démarche commune est la suivante : la première phase consiste donc à formaliser les textesafin qu'ils soient compréhensibles par la machine et utilisables par les algorithmesd'apprentissage. La catégorisation des documents est la deuxième phase, cette étape est bienentendu décisive car c'est elle qui va permettre ou non aux techniques d'apprentissage deproduire une bonne généralisation à partir des couples (Document, Classe).

Pour améliorer la performance des modèles, une évaluation de la qualité des classifieurs et lacomparaison des résultats fournis par les différents modèles est effectuée en fin de cycle. La démarche d'une approche standard de classification automatique de textes peut être résumée de la manière suivante :

- Eliminer les caractères de séparation, les signes de ponctuations, les mots vides, etc…
- Les termes restants sont tous des attributs.
- Un document devient un vecteur <terme, fréquence>.
- Entrainer le modèle de classification à partir des couples (Document, Classe).
- Évaluer les résultats du classifier.

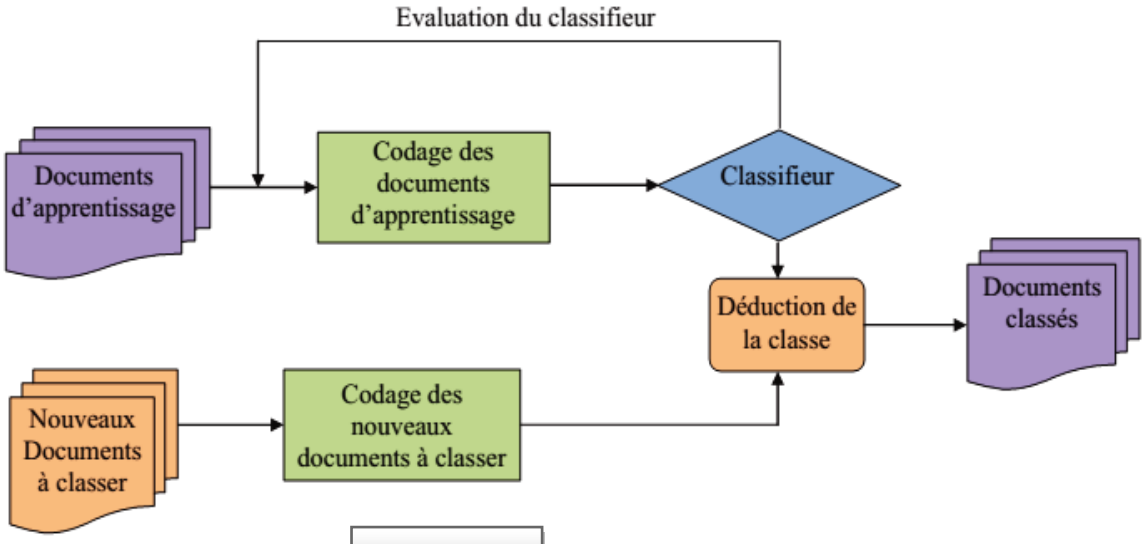


Figure 1

ii.3. Prétraitementsdes textes :

Le prétraitement des textes est une phase capitale du processus de classification, puisque laconnaissance imprécise de la population peut faire échouer l'opération. Après la première opération que doit effectuer un système de classification à savoir la reconnaissance des termes utilisés, nous devons filtrer le plus possible les informations inutiles des documents afin que les connaissances gardées soient aussi pertinentes qu'il sepeut. En effet dans les documents textuels de nombreux mots apportent peu (voir aucune) d'informations sur le document concerné. Les algorithmes dits de "Stop Words" s'occupentde les éliminer. Un autre traitement nommé "Stemming" permet également de simplifier lestextes tout en augmentant leurs caractères informatifs comme d'autres méthodes quiproposent de supprimer des mots de faible importance.

Toutes ces transformations et méthodes font partie de ce qu'on appelle le prétraitement. Plusieurs d'entre elles sont spécifiques à la langue des documents (on ne fait pas le mêmetype

de prétraitement pour des documents écrits en anglais qu'en français ou encore en arabe).

Le prétraitement est généralement effectué en six étapes séquentielles :

- La segmentation (isoler les ponctuations, les caractères de séparation, leschiffres).
- Suppression des mots fréquents
- Suppression des mots rares
- Le traitement morphologique
  - **Stemming** (regroupe sous un même terme (stem) les mots qui ont la même racine)
  - **lemmatisation** conserve, non pas les mots eux-mêmes, mais leur racine ou lemme(singulier/pluriel,conjugaisons,…)
- Le traitement syntaxique (traite les combinaisons et l'ordre des mots dans la phrase).
- Le traitement sémantique (extraire la signification des expressions et traiter la polysémie) .

ii.4. Prétraitement de texteArabe :

Tous les documents de texte arabe ont été prétraités selon les étapes suivantes :

- Conversion en encodage UTF-8.
- Suppression de traits d'union, des signes de ponctuation, des chiffres, des lettres non-arabes etdiacritiques.
- Retirer les Stop Words.
- Éliminer les mots rares (mots qui se produisent moins de cinq fois dans le jeu de données).
- Le vecteur norme espace modèle (VSM) a été utilisé pour représenter les textes arabes et TFIDF a été utilisé pour calculé le poids de termes.

ii.5. Codage des textes :

Un codage préalable du texte est nécessaire, comme pour l'image, le son, etc.,[Seb, 2002], car il n'existe pas actuellement de méthode d'apprentissagecapable de traiter directement des données non-structurées, ni dans la phase deconstruction du modèle, ni lors de son utilisation en classement. Pour la majorité des méthodes d'apprentissage, il faut transformer l'ensemble des textes en un tableau croisé "individus-variables" :

- L'**individu** est un texte (un document) $\boldsymbol{d_j}$, étiqueté lors de la phase d'apprentissage, et à classer dans la phase de prédiction.

- Les **variables** sont les descripteurs (les termes) $\boldsymbol{t_k}$qui sont extraits des donnéestextuelles.
- Le **contenu du tableau** (les éléments $\mathbf{w_{kj}}$), au croisement du texte j et du termek, représente le poids de ce terme k dans le document j.

ii.6. Choix de termes :

Dans la catégorisation de textes, on transforme le document $d_j$en un vecteur $\mathbf{d}_j$= (w_{1j}, w_{2j}, ..., $w_{|T|j}$), où T est l'ensemblede termes (ou descripteurs) qui apparaissent au moins une fois dans le corpus (ou la collection) d'apprentissage. Le poids w_{kj}correspond à la contribution des termest_kà la sémantique du texte d_j. Notons que la représentation par un vecteur entraîne uneperte d'information notamment celle relative à la position de mots dans la phrase.

ii.7. Représentation en " sac de mots " :

La représentation de textes la plus simple a été introduite dans le cadre du modèle vectoriel présenté ci-dessus, et porte le nom de « sac de mots ». L'idée est detransformer les textes en vecteurs dont chaque composante représente un mot.

Les composantes du vecteur sont une fonction de l'occurrence des mots dans letexte.

ii.8. Représentation des textes par des phrases :

Malgré la simplicité de l'utilisation de mots comme unité de représentation, certains auteurs proposent plutôt d'utiliser les phrases comme unité [Fuh, 1991]. Lesphrases sont plus informatives que les mots seuls, par exemple : « machine d'apprentissage »ou « world wide web» car les phrases ont l'avantage de conserver l'information relative à la position du mot dans la phrase (the problème of compositionnelsemantics).

Logiquement, une telle représentation doit obtenir de meilleurs résultats que ceux obtenus via les mots. Mais les expériences présentées ne sont pas concluantes car, si les qualités sémantiques sont conservées, les qualités statistiques sont largement dégradées : le grand nombre de combinaisons possibles entraîne des fréquences faibles ettrop aléatoires [jal,2003]. Néanmoins, proposent d'utiliser les phrases statistiques comme unités de représentation en opposition auxphrases « grammaticales » et obtiennent de bons résultats. Une phrase statistique estun ensemble de mots contigus (mais pas nécessairement ordonnés) qui apparaissentensembles mais qui ne respectent pas forcement les règles grammaticales. Afin de déterminer les phrases statistiques, [Car, 2001] utilisent des prétraitementstels que l'élimination des mots outils (stop words) et le stemming.

ii.8.1. Représentation des textes avec des racines lexicales et des lemmes:

La lemmatisation consiste à remplacer les verbes par leur forme infinitive, et les noms par leur forme au singulier. Un algorithme efficace, nommé TreeTagger [Schmid, 1994], a été développé pour les langues anglaise, française, allemande etitalienne.

L'extraction des stemmesrepose, quant à elle, sur des contrainteslinguistiques bienmoins fortes ; elle se base sur la morphologie flexionnelle mais aussi dérivationnelle. De ce fait, les algorithmes sont beaucoup plus simplistes et mécaniques que ceux permettant l'extraction des lemmes ; ils sont donc plus rapides ; maisleur précision et leur qualité sont naturellement inférieures.

ii.9. Méthodes basées sur les n-grammes :

Une autre approche pour coder les documents émerge : les n-grammes [Shannon, 1948]. On définit un n-gramme (n-gram) par est une séquence de n caractères : bi-grammes pour n=2, tri-grammes pour n=3, quadri-grammes pour n=4, etc…;

On n'a plus besoin de chercher lesdélimiteurs(les espaces ou les caractères de ponctuations) comme c'était le cas pour les mots. Quelques auteurs admettent les n-grammes comme une chaine non ordonnée de caractères; par exemple un tri-grammes peut être constitué du 2éme, 4éme et 1ér caractère, d'autres auteurs n'autorisent pas ce désordre. Pour notre cas, on va admettre qu'un n-grammes désignera une chaîne de n caractères consécutifs.

ii.10. Codage des termes:

Une fois choisies les composantes du vecteur représentant un texte j, il faut décider comment coder chaque coordonnée de son vecteur $\boldsymbol{d}_j$.

Il existe différentes méthodes pour calculer le poids $\boldsymbol{w}_{kj}$. Ces méthodes sont basées sur les deux observations suivantes :

- Plus le terme $t_k$est fréquent dans un document d_j, plus il est en rapport avec le sujet de ce document.
- Plus le terme $t_k$est fréquent dans une collection, moins il sera utilisé comme discriminant entre documents.

Soient $\#(t_k,d_j)$ le nombre d'occurrences du terme $t_k$dans le texte d_j, $|Tr|$ le nombre de documents du corpus d'apprentissage et $\#Tr(t_k)$ le nombre de documents de cet ensemble dans lesquels apparaît au moins une fois le terme t_k. Selon les deux

observations précédentes, un terme $t_k$se voit donc attribuer un poids d'autant plus fort qu'il apparaît souvent dans le document et rarement dans le corpus complet. La composante du vecteur est codée $f(\#(t_k,d_j))$, où la fonction f reste à déterminer. Deux approches triviales peuvent être utilisées. La première consiste à attribuer un poids égal à la fréquence du terme dans le document :

$w_{kj}= \#(t_k, d_j)$

et la deuxième approche consiste à associer une valeur booléenne

$$w_{kj}=1 \begin{cases} \text{Si } \#(t_k,d_j) > 1 \\ 0 \text{ Sinon} \end{cases}$$

ii.11. Codage TF × IDF :

Le codage TF × IDF a été introduit dans le cadre du modèle vectoriel et utilise une fonction de l'occurrence multipliée par $\frac{|Tr|}{\#Tr(t_k)}$une fonction de l'inverse du nombre de documents différents dans lequel un terme apparaît. Ce sigle provient de l'anglais et signifie « 'termfrequency'×'inverse document frequency' ».
Les termes caractérisant une classe apparaissent plusieurs fois dans les documents de cette classe, et moins, ou pas du tout, dans les autres. C'est pourquoi le codage TF × IDF [Seb, 2002] est défini comme suit :

$$\text{TF} \times \text{IDF}(t_k, d_j) = \#(t_k, d_j) \times \log$$

ii.12. Réduction de la dimension :

Un problème central pour l'approche statistique de la catégorisation de textes est la grande dimension de l'espace de représentation. Avec la représentation en sac de mots, chacun des mots d'un corpus est un descripteur potentiel ; or pour un corpus de taille raisonnable, ce nombre peut être de plusieurs dizaines de milliers. Pour beaucoup d'algorithmes d'apprentissage, il faut sélectionner un sous-ensemble de ces descripteurs. Sinon, deux problèmes se posent :

- le coût du traitement car le nombre des termes intervient dans l'expression de lacomplexité de l'algorithme ; plus ce nombre est élevé, plus le volume de calcul est important.
- la faible fréquence de certains termes : on ne peut pas construire des règles fiables à partir de quelques occurrences dans l'ensemble d'apprentissage.

On a observé que les mots les plus fréquents peuvent être supprimés : ils n'apportent pas d'information sur la catégorie d'un texte puisqu'ils sont présents partout. Demême, les

mots très rares, qui n'apparaissent qu'une ou deux fois sur un corpus, sont supprimés, car leurs faibles fréquences ne permettent pas de construire de règlesstables.

Cependant, même après la suppression de ces deux catégories de mots, le nombre de candidats reste encore élevé, et il faut utiliser une méthode statistique pour choisirles mots utiles pour discriminer entre documents pertinents et documents non pertinents, ou, plus généralement, entre les classes de documents.

a rappelé ce qui a été évoqué par [Seb, 2002] sur la réduction desdimensions qui peut être localement ou globalement :

- Réduction locale : Chaque classe est caractérisée par un profil composé d'un ensemble determes, et chaque texte sera représenté par une liste de termes dépendante de la catégorie.
- Réduction globale: Contrairement au cas précédent, un texte est représenté par une seuleliste de termes dans tous le corpus indépendamment des classes.

Conclusion :

Pour pouvoir appliquer les différents algorithmes d'apprentissage sur les documents de typetextuels, un ensemble de techniques ont été développé pour montrer comment l'informationtextuelle est habituellement prise en compte pour la représentation « informatique » de cesdocuments. Les différentes approches de représentation informatique de textes sont exposéesdans ce chapitre.
Ainsi avant la codification des documents, un ensemble d'opérations préliminaires doiventêtre faites pour épurer le texte de tous les mots inutiles et conserver seulement ceux qui sontporteurs d'informations et utiles pour le processus de classification. Mais malgré tous lesprétraitements appliqués sur le document, l'espace des descripteurs, qui peuvent être des n-grammes, des stems, des phrases, des concepts ou tout simplement des mots, reste très grandet très creux, d'où la nécessité d'une diminution préalable de cet espace.

Plusieurs techniques de sélection des descripteurs ou réduction de dimensionnalité sontproposées dans la littérature, une bonne partie de ces approches sont étalées dans ce chapitre.

Une fois la liste des descripteurs arrêtés, un degré d'importance ou poids est attribué à tous lestermes présents dans la représentation vectorielle ou probabiliste puisque chaque termepossède un certain nombre d'occurrences dans le document ou dans le corpus qui estdifférents aux autres.

Finalement on peut qualifier notre texte, par fichier « informatique » apte à être employé dansles différentes méthodes d'apprentissage automatique.

Chapitre 3 :

Catégorisation texte utilisant BPSO – KNN

III. Méthode BPSO / KNN Catégorisation :

III.1. Introduction :

Catégorisation des documents est un sujet important qui est au cœur de nombreuses applications qui exigent le raisonnement sur l'organisation et des documents texte, des pages Web, et ainsi de suite. Classification des documents est généralement obtenu en choisissant des caractéristiques appropriées (termes) et la construction de fréquence, et fréquenceinverse-documents (TFIDF) vecteur caractéristique. Dans ce processus, la sélection des attribut est un facteur clé dans la précision et l'efficacité des classifications qui en résultent.. Dans cechapitre, nous faisons démontrons une combinaison de PSO binaire « BPSO » et K voisin le plus proche « KNN » .

III.2. Optimisation par Essaim de Particules :

L'optimisation par essaim de particules (le PSO en anglais: Particle Swarm Optimazation) est une métaheuristique à base de population de solution. Elle a été proposée en 1995 par Kennedy et Eberhart [Ken, 1995]. L'algorithme PSO est inspiré du comportement social d'animaux évoluant en essaim, tels que les poissons qui se déplacent en bancs ou les oiseaux migrateurs. En effet, on peut observer chez ces animaux des dynamiques de déplacement relativement complexes, alors qu'individuellement chaque individu a une intelligence limitée et une connaissance seulement locale de sa situation dans l'essaim. L'intelligence globale de l'essaim est donc la conséquence directe des interactions locales entre les différentes particules de l'essaim. La performance du système entier est supérieure de la somme des performances de ses parties. Kennedy et Eberhart se sont inspirés de ces comportements sociaux pour créer l'algorithme PSO. Contrairement aux autres algorithmes évolutionnaires tel que l'algorithme génétique où la recherche de la solution optimale évolue par compétition entre les individus en utilisant des opérateurs de croisements et de mutations, le PSO utilise plutôt la coopération entre les individus. La méthode d'optimisation par essaim particulaire met en jeu un ensemble d'agents pour la résolution d'un problème donné. Cet ensemble est appelé essaim. L'essaim est composé d'un ensemble de membres, ces derniers sont appelés particules. Les particules de l'essaim représentent des solutions potentielles au problème traité. L'essaim de particules survole

l'espace de recherche, en quête de l'optimum global. Le déplacement de chaque particule est influencé par les trois composantes suivantes (Figure 3.):

- Une composante physique: la particule tend à suivre sa direction de déplacement courante.
- Une composante cognitive: la particule tend à se diriger vers le meilleur site par lequel elle est déjà passée;
- Une composante sociale: la particule tend à se diriger vers le meilleur site déjà atteint par ses voisines. Chaque particule i de l'essaim est définie par sa position $x_i = (x_{i1}; x_{i2};...; x_{iN})$ et sa vitesse de déplacement $v_{id}= (v_{i1}, v_{i2},..., v_{in})$dans un espace de recherche de dimension N. Cette particule garde en mémoire la meilleure position par laquelle elle est déjà passée et la meilleure position atteinte par toutes les particules de l'essaim, notées respectivement :$P_{bestid}= (P_{besti1}, P_{besti2},..., P_{bestid})$ et $G_{best}= (G_{best1}, G_{best2},..., G_{bestd})$. Le processus de recherche est basé sur deux règles :
- Chaque particule est dotée d'une mémoire qui lui permet de mémoriser la meilleure position par laquelle elle est déjà passée et elle a tendance à retourner vers cette position.
- Chaque particule est informée de la meilleure position connue au sein de son voisinage et elle a toujours tendance de se déplacer vers cette position. La particule i va se déplacer entre les itérations t et t+1, en fonction de sa vitesse et des deux meilleures positions qu'elle connaît (la sienne et celle de l'essaim) suivant les deux équations suivantes [Ken, 1995]:

$$v_{id}(t) = v_{id}(t-1) + c_1 r_1 (P_{bestid}(t-1) - x_{id}(t-1)) + c_2 r_2 (G_{bestd}(t-1) - x_{id}(t-1)) \quad \ldots(3)$$

$$x_{id}(t) = x_{id}(t-1) v_{id}(t) \quad \ldots\ldots(4)$$

Avec :

- $x_{id}(t)$, $x_{id}(t-1)$: la position de la particule *i* dans la dimension d aux temps t et t-1, respectivement.
- $v_{id}(t)$, $v_{id}(t-1)$: la vitesse de la particule i dans la dimension d aux temps t et t-1, respectivement.
- $P_{bestid}(t-1)$: la meilleure position obtenue par la particule i dans la dimension d au temps t-1.

  •$G_{bestd}(t-1)$: la meilleure position obtenue par l'essaim dans la dimension d au temps t-1.

c_1, c_2 : deux constantes qui représentent les coefficients d'accélération.

- r1, r2 : nombres aléatoires tirés de l'intervalle [0,1].

•v_{id} *(t-1),* c_1 r_1 $(P_{bestid}(t-1) - x_{id}(t-1))$, c_2 r_2 $(G_{bestd}(t-1) - x_{id}(t-1))$: représentent respectivement, les trois composantes citées au-dessus.

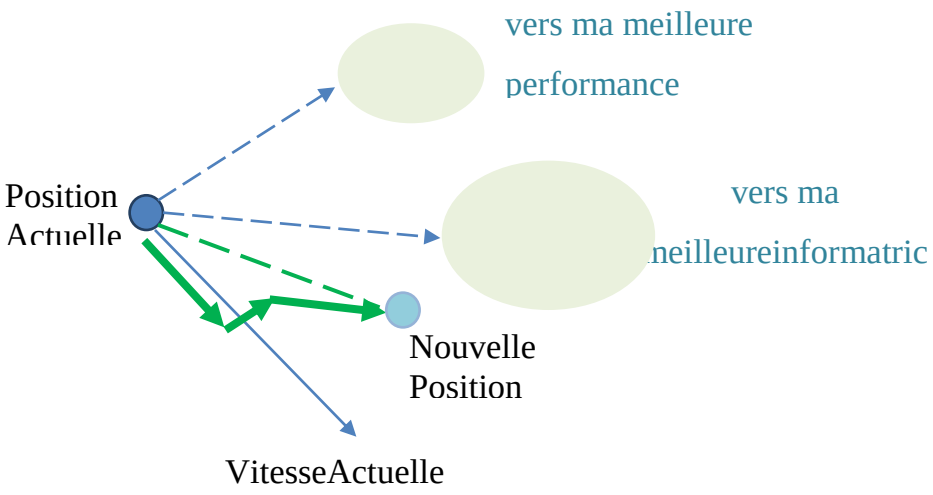


Figure 2

Afin d'estimer la qualité de la particule *i*, il est indispensable de calculer sa fonction «objectif» (aussi appelée fitness). La valeur de la fonction «objectif» de la particule x_{id} est notée $f(x_{id})$. Cette dernière est calculée en utilisant une fonction spéciale au problème traité.

Afin de mettre à jour les valeurs de x_{id}, P_{bestid} *et* G_{bestd}, leurs fitness sont calculées à chaque itération de l'algorithme. x_{id} est mise à jour selon l'équation (3). P_{bestid} et G_{bestd} sont mises à jour si les conditions (C1) et (C2) présentées ci-dessous sont vérifiées respectivement :

$f(x_{id})$ est meilleur que $f(P_{bestid})$……..(C1)

$f(P_{bestid})$ est meilleur que $f(G_{bestd})$……..(C2)

Algorithme 1 : L'optimisation par essaim de particules

Début

Initialiser les paramètres et la taille S de l'essaim;

Initialiser les vitesses et les positions aléatoires des particules dans chaque dimension de l'espace de recherche;

Pour chaque particule, $P_{bestid} = x_{id}$;

Calculer f (x_{id}) de chaque particule;

Calculer G_{bestd} ; // la meilleure P_{bestid}

Tant que (la condition d'arrêt n'est pas vérifiée) **faire**

Pour (i allant de 1 à S) **faire**

Calculer la nouvelle vitesse à l'aide de l'équation (3) ;

Trouver la nouvelleposition à l'aide de l'équation (4) ;

Calculer f (x_{id}) de chaque particule ;

Si (f (x_{id}) est meilleur que f (P_{bestid})) **alors**

P_{bestid}= x_{id};

Si (f (P_{bestid}) est meilleur que f (G_{bestd})) **alors**

G_{bestd} = P_{bestid};

Fin pour

Fin tant que Afficher la meilleure solution trouvée G_{bestd} ;

Fin

I.1. Topologies des voisinages :

Il y a 3 topologies principales de voisinage utilisées dans PSO (voir la figure): cercle, rayon et étoile. Le choix de la topologie de voisinage détermine quel $\mathbf{G_{bestd}}$employer pour l'individu X_i. Dans la topologie de cercle, chaque individu est connecté socialement à ses k voisins topologiques les plus proches (G_{bestd}de X_i = mieux résultat individuel parmi ses k voisins plus proches, k égale typiquement 2). La topologie de rayon, isole efficacement les individus les uns des autres, car l'information doit être communiquée par un individu focal.La topologie d'étoile est lameilleure topologie connue ; ici chaque individu est relié à chaque autre individu ($\mathbf{G_{bestd}}$de X_i = mieux différents résultats dans la population).

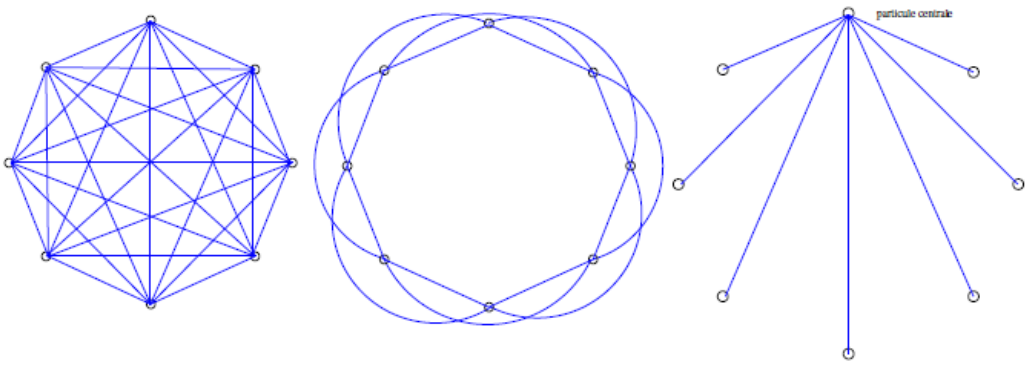


Figure 3

I.2. Méthode de Classification Texte Arabe avec BPSO / KNN:

I.2.1. L'optimisation par essaim de particules binaire BPSO :

La première version de l'algorithme BPSO (l'algorithme standard binaire de l'optimisation par essaim de particules) a été proposée en 1997 par Kennedy et Eberhart [Ken, 1997]. Dans l'algorithme BPSO (de l'anglais:BinaryParticleSwarmOptimization), la position de la particule i est représentée par un ensemble de bit. La vitesse v_{id}de la particule *i*est calculée à partir de l'équation (3). v_{id}est un ensemble de nombres réels qui doivent être transformés en un ensemble de probabilités en utilisant la fonction sigmoïde comme suite :

$$S(V_{id}) = \frac{1}{1+e^{-V_{id}}} \quad \text{…….. (5)}$$

Où S (V_{id}) représente la probabilité du bit X_{id}de prendre la valeur 1.En se basant sur la vitesse V_{id}de la particule i et sur la valeur de la probabilité $S(V_{id})$, la position de la particule X_{id}est calculée comme suite :

$$X_{id} = \begin{cases} 1 \text{ Si } r < S(V_{id}) \\ 0 \text{ Sinon} \end{cases} \quad \text{…….. (6)}$$

I.2.2. K plus proches voisins – kPPV (KNN) :

La méthode des k plus proches voisins ou The k-NN classification (k-Nearest Neighbors) est un classifieur à base d'instances qui fait partie des méthodes géométriques utilisant des mesures de distance.

k-NN est un algorithme classique d'apprentissage qui a été longtemps à la base des algorithmes de catégorisation des documents, elle a été employée avec succès dans de domaine de classification et a engendré toute une famille de classifieurs connus sous le nomade classifieurs paresseux (lazy learns). Dans ces systèmes, Le classifieur calcule lassimilarité du nouveau texte à catégoriser avec l'ensemble des autres exemples du corpusd'apprentissage, dont les catégories sont déjà connues, puis il sélectionne les k documents lesplus proches du document à classer. Ensuite, pour affecter la catégorie, les relations entre cesk documents et les catégories sont évaluées et un score est calculé par catégorie afin d'évaluerla pertinence de la catégorie au document. La catégorie (ou les catégories) ayant le score leplus élevé (celle qui contient le plus de textes voisins) est affectée au document [Yan,2001]. Voici son algorithme général :

Algorithme 2 : K plus proches voisins

Paramètres : le nombre k de voisins
Données : un ensemble d'exemples classés (document, classe)
Entrée : un nouveau document D
1. déterminer les k plus proches documents de D
2. Sélectionner la classe majoritaire C des classes de ces k exemples
Sortie : la classe de D est C

I.2.3. Hybridation BPSO/KNN :

Dans cette section, nous décrivons la méthode de Catégorisation, qui est une hybridation BPSO / KNN, étape par étape , tout cela fait précéder par une étape de prétraitement de texte, décrit dans la section , dans lequel un total de N termes (caractéristiques) sont prédéterminée à partir de la collection (Corpus) de documents.

Étape (1): Créer une population de particules sur N dimensions dans l'espace des classes.

Chaque particule est représentée par trois vecteurs:

- la position actuelle de particule (X_i).
- lameilleure position précédente particule (P_i)
- et sa vitesse (V_i).

X_i est initialisé avecvaleurs binaires aléatoires où 1 signifie que le classe correspondant est sélectionné et 0indique l'inverse. P_i est initialisé avec une copie de X_i. (Suite à l'évaluation
de chaque particule dans l'essaim, la G_{best} Global est initialisée avec l'index de la particule avec la meilleure valeur de fitness).

Étape (2): Pour chaque particule:

- Évaluer fitness utilisant KNN classifier (voir ci-dessous).
- Mise à jour la meilleure position personnellede particules.

Étape (3): Mise à jour de G_{best}Global.

Étape (4):mise à jour toutes lesvélocitéset les positions des particules de la population selon l'approche standard dans BPSO .

$$v_{id}(t) = \omega\ v_{id}\ (t\text{-}1) + c_1\ r_1\ (P_{bestid}\ (t\text{-}1) - x_{id}\ (t\text{-}1)) + c_2\ r_2\ (G_{besti}\ (t\text{-}1) - x_{id}\ (t\text{-}1)) \ldots. \quad (7)$$

Où ω est le coefficient d'inertie

La probabilité de changement de bit est déterminée par la formule(6) au-dessus.

Étape (5): Terminer si le critère de terminaison est satisfaite, la sortie sélectionné
La catégorie (représenté par la meilleure particule Globale courante), sinon passez à l'étape (2).

La fitness d'une particule est calculée selon la formule suivante :

$$Fitness = (\propto * Acc) + (\beta * (\frac{N-T}{N})) \quad \ldots\ldots\ldots\ldots\ldots\ (8)$$

Où

- *Acc* est la précision de la classification de la particule trouvé avec KNN (voir ci-dessous).
- α et β sont deux paramètres utilisés pour équilibrer entre la précision de classification

et la taille de catégorie (sélectionné par les particules),

α est dans l'intervalle [0,1] et β= α-1

- N est le nombre total de catégories.
- T est la taillede la catégorie sélectionnéepar particule.

La précision de classification d'une particule (P) est calculée en utilisant procéduresuivant :

- Filtrer la catégoriesélectionnée par particule.
- Set C = 0.
- Pour chaque instance de corpus d'apprentissage (ce qui est au cours de la phase d'apprentissage).
- Calculez la distance euclidienne de l'instance en cours à toutes les instances de corpus d'apprentissage.
- Classez l'instance en cours en fonction de ses K voisins les plus proches dans Corpus d'apprentissage.
- Si la classification trouvée correspond à la classification connue de l'instance, augmenter C par 1.
- Enfin, la précision de la classification de P est enregistrée comme C divisé par le nombre totalcorpus d'apprentissage.

I.2.4. Paramètres BPSO / KNN :

Les expériences de BPSO / KNN concentrées sur des paramètres suivants :

- Coefficient d'inertie :entre 0.9 et 1.2.
- Nombre de générations : 100 générations.
- K paramètre pour classifierKNN : K=3 classe.
- Taille d'essaim : 30
- mots rares: les mots qui apparaissent moins de 5 fois dans le corpus d'apprentissage ont été enlevés.

I.3. Conclusion :

Dans ce chapitre, nous concluonsque la catégorisation de texte baséesur BPSO KNNa la simplicité et des paramètres limités et une précision élevé comparons par d'autre méthode de catégorisation.

Chapitre 4 :

Conception et Implémentation

IV. Conception et Implémentation

IV.1. Introduction :

Dans la première partie de ce chapitre, nous allons faire la conception de notre application en utilisant le langage UML .cependant notre application n'est pas d'une grande complexité afin d'utiliser la totalité des diagrammes d'UML, nous n'utiliserons que les diagrammes que nous trouverons nécessaires.

La deuxième partie de chapitre consacré à la présentation de la mise en œuvre de notre application, nous commençons tout d'abord par une présentation du langage de programmation choisi, ainsi que les motivations de ce choix. Ensuite nous mentions les détails de classification des documents textuelles . On se termine ce chapitre par une synthèse de nos résultats obtenus.

IV.2. Présentation de langage UML:

UML (UnifiedModelingLanguage, que l'on peut traduire par "langage de modélisation unifié") est une notation permettant de modéliser un problème de façon standard. Ce langage est né de la fusion de plusieurs méthodes existant auparavant, est devenu désormais la référence en terme de modélisation objet, à un tel point que sa connaissance est souvent nécessaire pour obtenir un poste de développeur objet [GRA, 98]. UML permet à des programmeurs utilisant des langages de programmation différents de parler la même langue. De plus, UML est un langage simple et intuitif. Elle est constituée de 13 diagrammes qui utilisent des symboles sémantiques bien précis. Ils permettent de concevoir une modélisation fiable limitant ainsi beaucoup les grosses erreurs de programmation (erreurs de conception).

IV.3. La conception de l'application:

Le principe général de notre travail est la classification des messagesenvoyés par l'utilisateur "mineur" via Facebook, afin de protégé ce dernierde dangers rencontrée en Facebook, notre application se déroule en deux phases. La première phase est une phase.

IV.4. Les diagrammes utilisés :

IV.4.1. Le diagramme des cas d'utilisation :

L'ensemble des cas d'utilisation décrit exhaustivement les exigences fonctionnelles et techniques du système. Chaque cas d'utilisation correspond donc à une fonction métier du système, selon le point de vue d'un de ses acteurs.

Dans notre application l'utilisateur doit choisir leur profil et ensuite s'authentifier à leur compte, si l'utilisateur est un administrateur il peut ajouter une nouvelle classe.

Les cas d'utilisation de notre système sont représentés par la figure suivante :

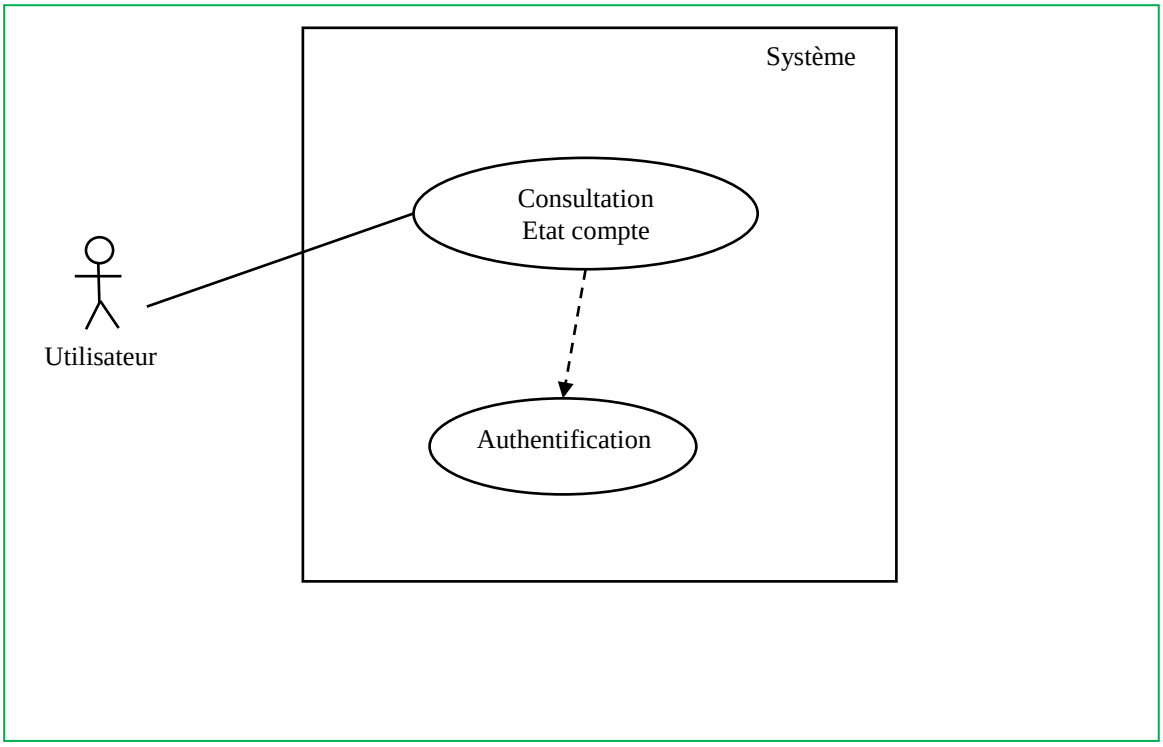
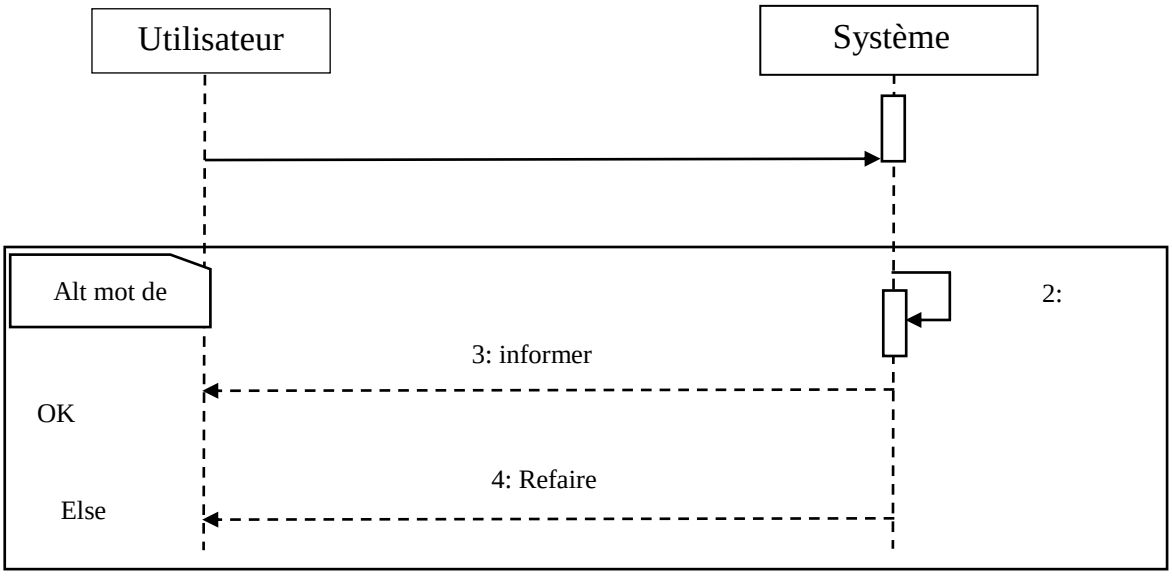


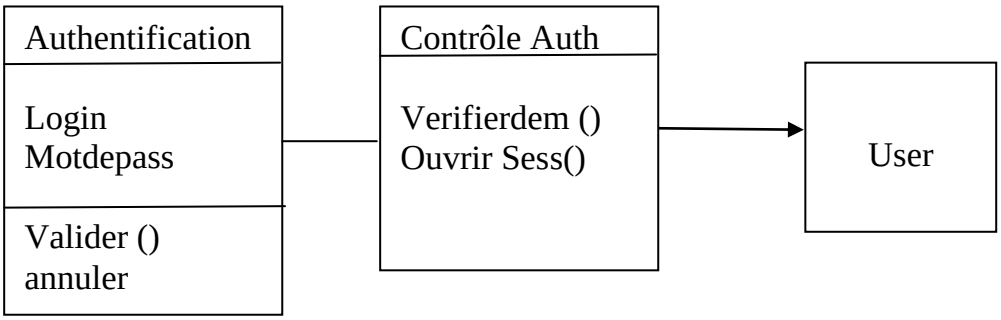
Figure 4

IV.4.2. Diagramme de séquences :

les diagrammes de séquences permettent de représenter des collaborations entre objets selon un point de vue temporel et peuvent servir à illustrer des cas d'utilisation. Donc, on y met l'accent sur la chronologie des envois de messages.

- **Diagramme de séquences «Consultation Compte» :**





IV.5. La réalisation :

IV.5.1. Notre approche :

Nous présentons dans cette partie l'approche que nous avons retenue pour protéger les utilisateurs "mineurs" de ***Facebook***de deux dangers majeurs : la classification de tous les messages envoyés par les mineursdans des classes bien définies. Nous avons choisi *Facebook* parce qu'il est actuellement le réseau social le plus populaire comparé à d'autres réseauxsociaux.

Nous avons créé une application de classification Son architecture générale est représentée par la figure suivante :

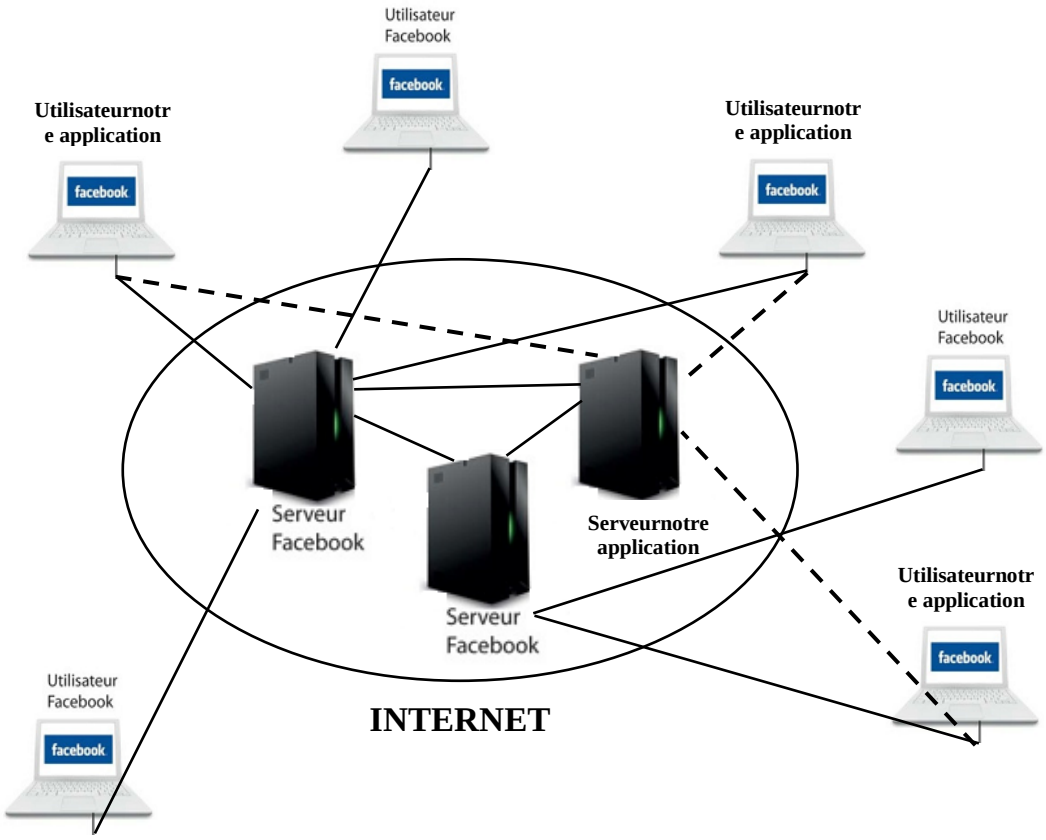
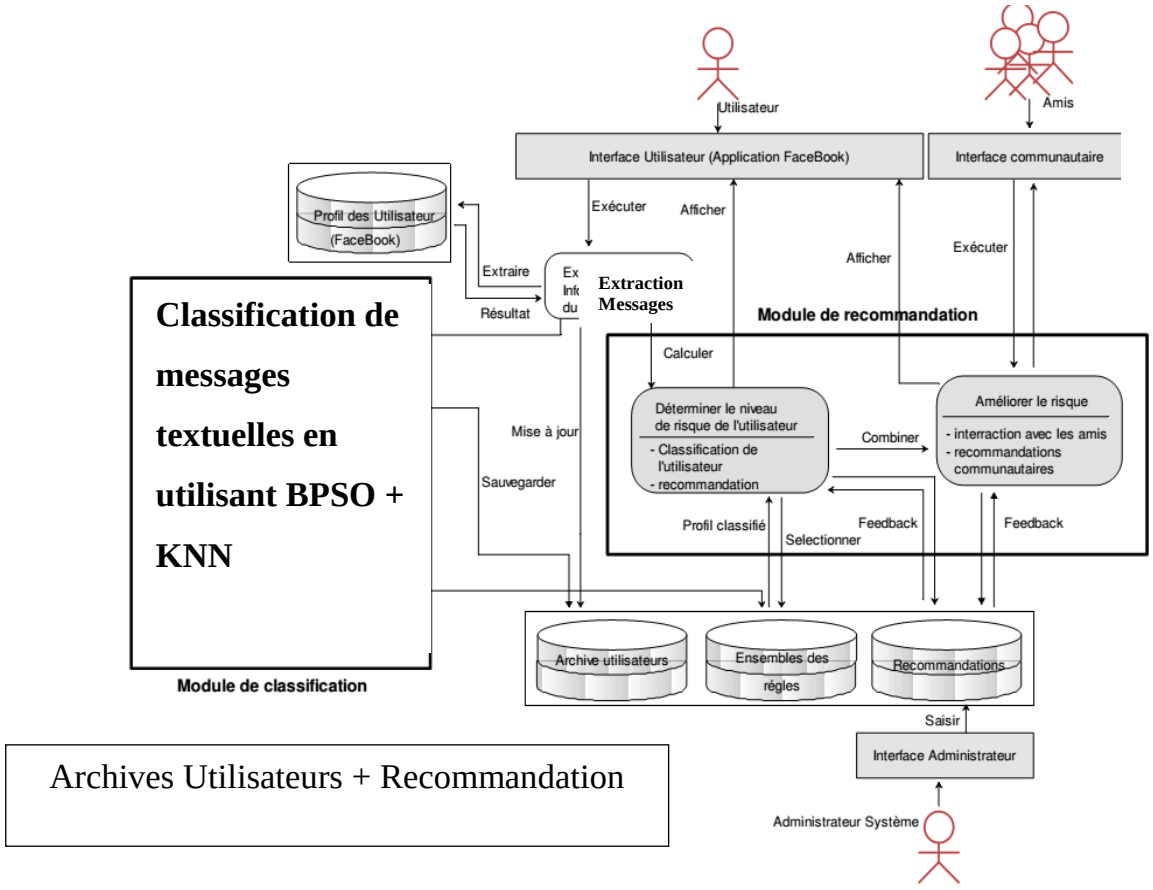


Figure 5

Ils ont été développés avec les langages *PHP* en utilisant l'*API*(*Application Programming Interface*) et la plateforme *SDK 5.0* (*Software Développement Kit*)que ***Facebook***met à la disposition des développeurs. Cette dernière permet l'interaction entre les utilisateurs et le serveur *Facebook*. Notre système s'exécute sur un serveur *PHP*qui est connecté à un serveur « *.NET* ».

Notre application est constituée de deux modules : le module de classification etle module de recommandation. Le fonctionnement et l'architecture de ces 2 modules sont détaillés à La figure ci-dessous :



Recommandation à l'utilisateur de tous les messages classés interdits

Figure 6

IV.5.2. Interface Utilisateur :

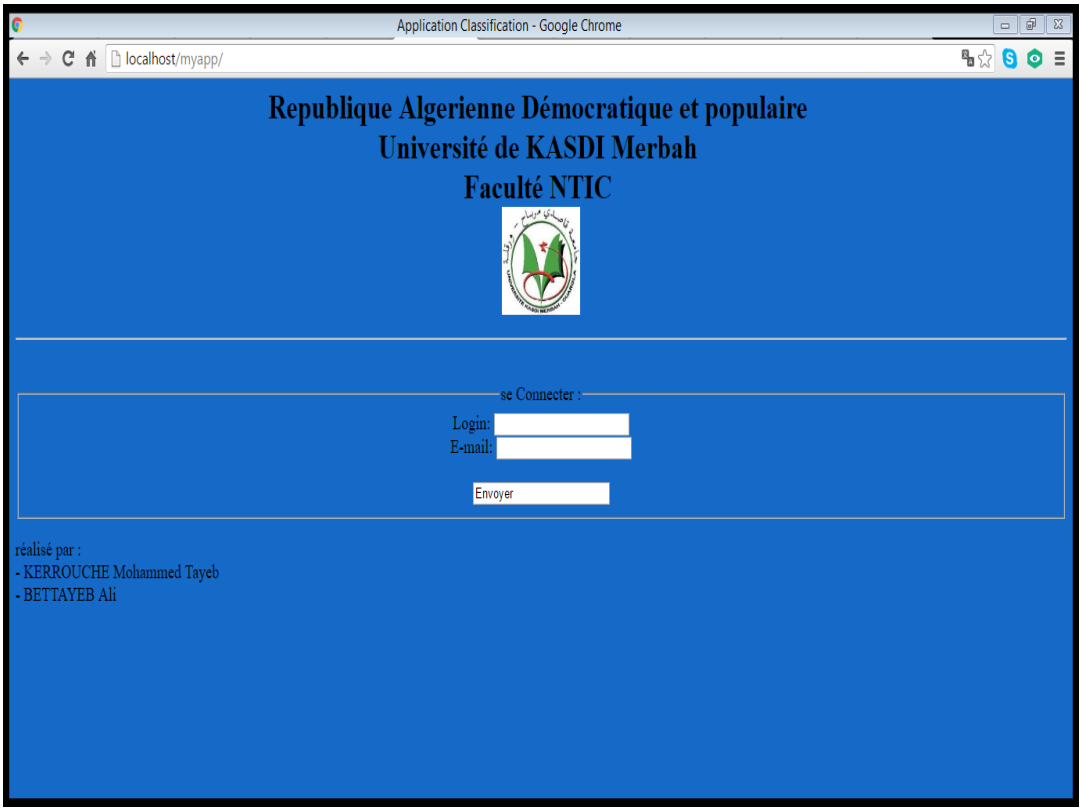


Figure 7

IV.5.3. Interface Extraction Message :

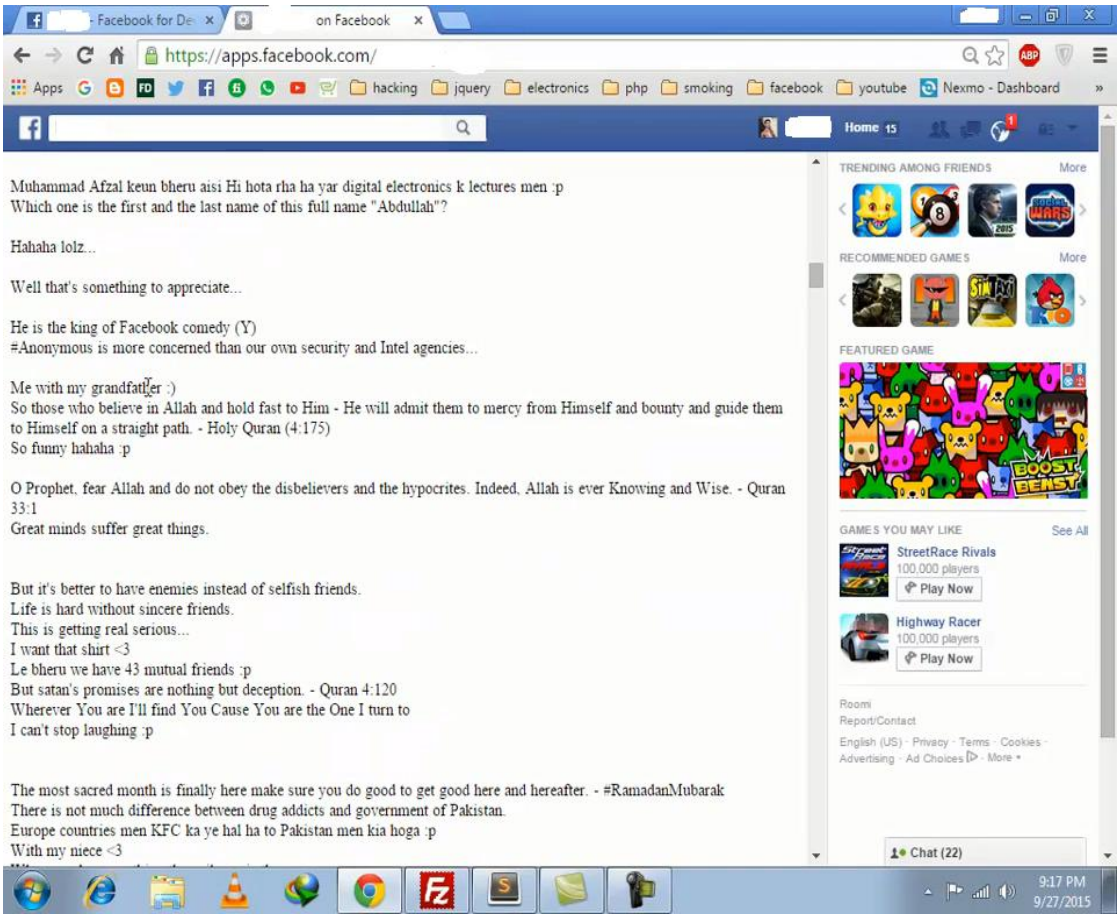


Figure 8

IV.5.4. Conclusion Générale :

Malgré les efforts et les travaux intensifs réalisés dans le domaine de la protection dans le Web et spécialement dans les réseaux sociaux, aucune méthode n'est jugé fiable à 100%, Mais au fur et à mesure les autres essayent d'améliorer les scores pour de meilleurs résultats. Pour notre travail de protection les mineurs, et notre approche de classification des messages textuelles envoyés par l'utilisateur, Nouschoisissons une méthode d'extraction des messages en utilisant le API Facebook. un classificateur BPSO hybridé par KNN pour catégorisé ces messages, et un corpus « dataset » d'apprentissage contient des milliers d'articles de crimes, racisme, mots pornographie, terrorisme, magique, violence…notre classificateur est justifié par leur efficacité dans les problèmes de classification et par leur principal avantage "apprentissage" qui permet de développer des systèmes dynamiques et évolutifs. .notreapplication peut être intégré dans le plateforme Facebook, et utilisé API de Facebook.

Bibliographie :

[Car, 2001] : M.F.Caropreso, S.Matwin, F.Sebastiani , A learner-independent

[GRA, 1998] : Grady Booch, James Rumbaugh, Unified Modeling Language User Guide , Article, 1998.

[jal,2003] : R.JALAM, Apprentissage automatique et catégorisation de textes multilingues, 2003.

[Ken, 1995] : J. Kennedy and R. Eberhart, Particle Swarm Optimization, 1995.

[Ken, 1997] : J. Kennedy, R.C. Eberhart. A discrete binary version of the particle swarm algorithm, 1997.

[Schmid, 1994] :H.Schmid , Probabilistic part-of-speech tagging using decision trees .

[Seb, 2002] :F.Sebastiani , Machine learning in automated text categorization , 2002.

[Sha, 1948] : C.Shannon ,The Mathematical Theory of Communication , 1948

[Yan, 2001] Y.Yang , Problem-based Learning on the World Wide Web in an Undergraduate evaluation of the usefulness of statistical phrases for automated text categorization ,2001.

Kinesiology Class: an Integrative Approach to Education , 2001

Site Web :

1. [Web, 01] :http://netintelligenz.net/post/88275150188/pourquoi-les-jeunes-pr .

[Web, 02] : https://www.cjc-ccm.gc.ca/.../JTAC-Ssc-Report-to-JTAC-2010-01-29

Résumé

Le nombre d'usagers des réseaux sociaux augmente annuellement à une très grande vitesse, Des milliers de comptes usagés sont créés quotidiennement. Un nombre incalculable de données lues et partagées par les différents comptes. Ceci provoque des dangers illimités sur beaucoup d'utilisateurs de ces réseaux sociaux surtout les mineurs. Il est donc crucial de sensibiliser ces utilisateurs aux dangers potentiels qui les guettent. Nous présentons une application pour La protection des mineurs contre les dangers de Facebook, notre approche est basées sur des techniques bio-inspirés "BPSO" hybridé par "KNN" pour la catégorisation des messages textuels envoyés par le mineur. Si la catégorie est interdite par le parent, notre application notifier ce parent par ces messages.

Mots clés : Facebook, protection de mineurs, BPSO, KNN, catégorisation, messages.

Abstract

The number of users of social networks increases annually at a very high speed, Thousands of used accounts is created daily. Countless read and shared by different accounts data. This causes unlimited dangers on many users of these social networks especially minors. It is therefore crucial to educate these users to potential dangers that the guettent. we present an application for the protection of minors against the dangers of Facebook, we base on bio-inspired technology "BPSO" hybridized by "KNN " for the categorization of text messages sent by the minor on Facebook. If the category is forbidden by the parent, our application notify that parent by these messages.

Key words : Facebook, Minor protection, BPSO, KNN, categorization, messages.

ملخص

يزيد عدد مستخدمي الشبكات الاجتماعية سنويا بسرعة عالية جدا، يتم إنشاء آلاف الحسابات يوميا. عدد لا يحصى من البيانات (صور – رسائل نصية ...) يتم تقاسمها بين حسابات مختلفة. هذا يسبب مخاطر غير محدودة للعديد من المستخدمين لهذه الشبكات الاجتماعية خصوصا القصر. ولذلك فمن الأهمية بمكان لفت انتباه هؤلاء المستخدمين إلى الأخطار المحتملة للشبكات الاجتماعية على القصر خصوصا الفيس بوك، نحن نبني مقاربتنا للحماية على أسس مستوحاة من تحسين سرب الجسيمات الثنائية الحيوي "BPSO" تهجين" KNN" لتصنيف الرسائل النصية المرسلة من قبل القاصر. إذا كانت الفئة ممنوعة من قبل الوالدين،فإن برنامجنا يرسل إخطارا إلى الوالد.

كلمات مفتاحية :الفيسبوك،حماية الأحداث، تصنيف، الرسائل..KNN، BPSO ،