Apprentissage bayésien incrémental pour la détermination de l'âge et du genre d'utilisateurs de plateformes du web social

Jugurtha Aït-Hamlat[1, 2]
(1) Semantiweb, 2, rue Paul Vaillant Couturier, 92300 Levallois-Perret, France
(2) ERTIM, 2 rue de Lille, 75007 Paris, France
jugurtha.ah@gmail.com

RÉSUMÉ

Les méthodes de classification textuelles basées sur l'apprentissage automatique ont l'avantage, en plus d'être robustes, de fournir des résultats satisfaisants, sous réserve de disposer d'une base d'entraînement de qualité et en quantité suffisante. Les corpus d'apprentissage étant coûteux à construire, leur carence à grande échelle se révèle être l'une des principales causes d'erreurs. Dans un contexte industriel à forte volumétrie de données, nous présentons une approche de prédiction des deux plus importants indicateurs socio-démographiques « âge » et « genre » appliquée à des utilisateurs de forums, blogs et réseaux sociaux et ce, à partir de leurs seules productions textuelles. Le modèle bayésien multinomial est construit à partir d'un processus d'apprentissage incrémental et itératif sur une vaste base d'entraînement semi-supervisée. Le caractère incrémental permet de s'affranchir des contraintes de volumétrie. L'aspect itératif a pour objectif d'affiner le modèle et d'augmenter ainsi les niveaux de rappel & précision.

ABSTRACT

UGC text-based age & gender author profiling through incrementally semi-supervised bayesian learning

Text classification tasks based on machine learning, beside their robusteness, provide satisfactory results as long as it relies on sufficiently large and accurate training dataset. Supervised training data being expensive to build, their insufficiency appears to be one of the main causes of errors. In an industrial widespread data context, we present a predictive approach applied to the most interesting socio-demographic features (Age & Gender) of web users generated contents platforms (forums, blogs and social media). Through an incrementally iterative learning process, a multinomial bayesian model is built from a large semi-supervised training data. The incremental nature allows to overcome volume constraints while iterative aspect aims to refine the model features, improving tagging recall & precision levels.

MOTS-CLÉS : Classification de documents, Bayésien naïf multinomial, Apprentissage semi-supervisé, Profilage d'utilisateurs.

KEYWORDS: Text classification, Multinomial naive bayes, Semi-supervised machine learning, Author profiling.

1 Introduction

1.1 Contexte général

Le champ de recherche dans lequel s'inscrit ce projet a été déterminé par le contexte applicatif de l'entreprise Semantiweb [1]. L'activité de l'entreprise, axée sur le décodage marketing de diverses données d'utilisateurs, exploite les contenus issus des plateformes du web 2.0 [2]. Le travail ici présenté, appréhendé en tant que problématique de fouille sémantique de textes, a pour perspective la mise au point d'une méthode de profilage des auteurs issus du web social.

Nous entendons par « profilage des auteurs » l'estimation des deux indicateurs socio-démographiques « âge » et « genre » de ces utilisateurs et ce, à partir de leurs seules productions textuelles. Les résultats de cette classification sont pris en compte dans des études marketings afin de quantifier les tendances des internautes vis-à-vis de problématiques en lien avec les comportements de consommation, la perception de la cherté, les demandes de conseil ou les avis sur les produits et les services entre autres.

La méthode présentée ici décrit une approche de classification identique pour les deux indicateurs « âge » et « genre ». Cela dit, cette approche est utilisée séparément pour les deux indicateurs. En conséquence, il n'existe pas d'interdépendance dans le processus de classification de ces deux indicateurs.

1.2 Pourquoi les contenus générés par les utilisateurs ?

Les nouvelles technologies du web ont fait émerger une diversité de supports d'interaction qui aujourd'hui a révolutionné les pratiques de production et d'accès aux contenus. Le « post » émis par l'utilisateur y représente le mode d'expression le plus prolifique. Il se manifeste aussi bien dans les fils de discussions, les commentaires de diverses plateformes (vidéos, médias etc.), les articles de blogs, les posts des réseaux sociaux, les tweets ou encore les avis et les évaluations de produits ou de services.

Depuis une quinzaine d'année, les forums de discussion représentent l'un des espaces d'interaction asynchrone les plus utilisés sur le web. Avant d'être rejoints par les blogs et les réseaux sociaux, ces plate-formes permettent aux internautes d'échanger sur des thématiques de la vie quotidienne aussi variées que la famille, le travail, le couple, la santé etc. tout en préservant un certain anonymat, contrairement aux réseaux sociaux, dont le lien avec les autres utilisateurs est plus formalisé.

Au delà des aspects liés à l'audience, l'expansion de ces modes d'échange a favorisé la construction d'une représentation et/ou d'une expertise collective(s) à laquelle adhère la "communauté" des forumeurs, blogueurs, youtubeurs, twittos et facebookeurs. Les contributeurs se trouvent ainsi être les garants d'un lien et d'un savoir pratique à travers la confrontation des expériences et l'échange d'informations. Les forums et les blogs sont devenus de ce point de vue une sorte de référence légitime, un lieu spontané de questionnement, d'échange d'expériences, de partage de solutions et d'astuces mais aussi l'endroit où l'on parle de soi, de ses préoccupations quotidiennes et sociales. Au regard des annonceurs et des industriels, ces informations représentent une mine d'or au vu de ce

1. `www.semantiweb.fr`
2. Ici les plateformes de forums, de blogs et des réseaux sociaux. Il s'agit en l'occurrence des « contenus générés par les utilisateurs » (en. UGC).

qu'elles recèlent en matière de perceptions, d'attentes, de problèmes ou d'avis sur des services, des marques ou des produits de consommation.

La démocratisation des espaces d'expression grâce aux plateformes du web 2.0 engendre, au-delà du besoin d'innovation dans le domaine de la gestion des contenus, la nécessité d'exploiter ces données au travers de modèles innovants. La masse de données textuelles non-structurées ou semi-structurées que représentent aujourd'hui les flux générés par le web social est d'une importance telle qu'il est nécessaire de recourir à des méthodologies d'analyse performantes.

2 Etat de l'art

Le travail présenté ici s'appuie sur les apports de la linguistique de corpus et du traitement automatique des langues (TAL). Le recours aux outils de traitement de corpus permet de rendre observables des régularités qui ne l'auraient pas été sans le recours à des corpus constitués. Le décodage de la sémantique interprétative (Rastier, 2009) apporte dans ce cadre un éclairage sur le plan théorique.

L'intérêt suscité par l'étude des nouveaux modes d'interaction sur le web recèle des enjeux économiques, juridiques et sécuritaires d'une telle importance qu'ils font émerger de nouveaux besoins en efficacité et en robustesse quant à leur exploitation. La diversification des perspectives de recherches en TAL & REI [3] a de ce point de vue permis d'élargir les champs d'applications, allant de la e-réputation à la veille économique et stratégique, de la classification de documents à la reconnaissance des entités nommées, en passant par la fouille d'opinion ou l'analyse des sentiments (Eensoo et Valette, 2015).

Parmi ces perspectives, le profilage des auteurs suscite un intérêt grandissant ces récentes dernières années. À l'instar de la Reconnaissance des Entités Nommée (Nadeau & Sekine, 2007), cette tâche de fouille de textes est considérée aujourd'hui comme un champ d'application à part entière, donnant lieu à des évenements/concours à l'image des actes de conférences PAN (Rangel *et al.*, 2013). Les principaux axes d'exploration portent sur les indicateurs socio-demographiques comme l'âge, le genre mais aussi la domiciliation, la situation maritale ou socio-professionnelle. D'autres indicateurs sont également explorés à savoir le niveau d'éducation, les traits de personnalité ou la nativité par rapport à la langue.

Les données utilisées dans ce cadre proviennent de diverses sources, allant d'articles de blogs (Mukherjee & Liu, 2010), (Santosh *et al.*, 2013) aux e-mails (Estival *et al.*, 2007), de twitter (Burger *et al.*, 2011) ou de plateformes de tchat (Lin, 2007) aux divers médias sociaux (Marquardt *et al.*, 2014), (Nguyen *et al.*, 2014) en passant par des avis (Rangel *et al.*, 2015), (Otterbacher, 2010) ou des commentaires des utilisateurs de forums. Par ailleurs, si la répartition des classes du genre est communément admise étant donné son caractère binaire et discret [masculin, féminin], il en est différemment de l'âge. En effet, la granularité des classes d'âge se présente diversement selon les cas, à savoir à l'unité ou selon des tranches d'âges discretes diversement distribuées en nombre et en intervalles.

Par ailleurs, les méthodes de détermination des profils reposent pour la plupart sur un apprentissage supervisé (Rangel *et al.*, 2015). Quant aux modèles utilisés, ils sont communément à base d'algorithmes statistiques, parmi lesquels les modèles bayésiens (Chen *et al.*, 2009), de régression (Nguyen *et al.*, 2011), de clusturing ou des arbres de décision.

3. Recherche et Extraction d'Informations

Enfin, les résultats obtenus reposent, selon les travaux et les méthodes utilisées, soit sur des critères lexicaux (vocabulaire spécifique, variation) ou grammaticaux (constructions syntaxiques, accords, POS) ou encore orthographiques (qualité, ponctuation, majuscule) ou stylistiques (Weren *et al.*, 2014) (expressions prototypiques, registre de la langue, marqueurs de sentiments etc.) avec des différences de pré-traitements et de segmentation (tokens, Ngram, collocations) (Santosh *et al.*, 2013).

3 Corpus

La délimitation du corpus de travail a relevé d'un échantillonnage représentatif à partir d'une vaste masse des données. En effet, le contexte industriel dans lequel ce travail est mené se caractérise par la présence d'une forte volumétrie de données, hétérogènes et évolutives. Ces contraintes sont ici prises en compte afin de mettre en place un système de profilage des auteurs générique, robuste et dynamique.

3.1 Les posts :

Les données brutes sont collectées à partir d'une centaine de sources, essentiellement des forums de discussion mais aussi des plateformes de blogs (regroupant des centaines de milliers d'articles et leurs commentaires), ainsi qu'une centaine de pages publiques de Facebook et de Twitter. (cf. tableau 1)

Le « post » à proprement parler représente l'unité d'analyse de notre corpus. Il correspond à une contribution mise en ligne par un utilisateur sur une plateforme donnée, à un moment T. A côté de l'élément textuel, le post comprend généralement un pseudonyme, une date de mise en ligne, une URL et facultativement une ancre, un titre et un topique.

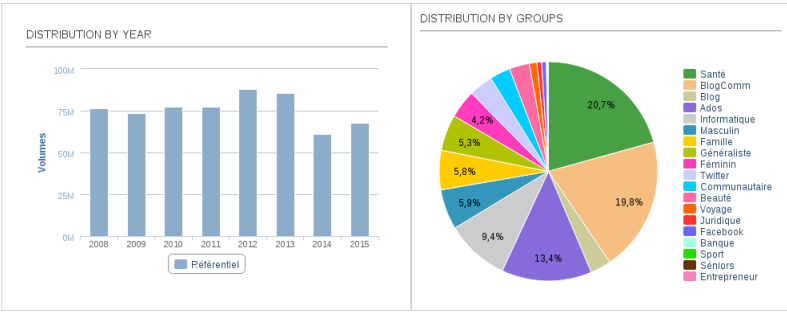

FIGURE 1 – Répartition des posts par années et par groupes (Interface Data Semantiweb)

3.2 Les auteurs :

À un auteur correspond un utilisateur ayant publié un ou plusieurs posts via une source sur laquelle il est préalablement inscrit. Il est identifié distinctement grâce à son pseudonyme et appartient à la seule source dont il est membre. Dans l'optique de la classification des profils, le lien entre

le post et l'utilisateur est indispensable et permet, grâce à la classification séparée des différents posts d'un utilisateur donné, d'estimer la classe la plus probable liée à cet utilisateur. Hormis les « hyper contributeurs » auxquels il convient de porter une attention particulière, il existe des types d'utilisateurs à exclure de ce processus en raison de problème lié à l'unicité de leur pseudonymes. Il s'agit en l'occurrence d'utilisateurs supprimés, administrateurs, modérateurs, non-membres, invités et autres anonymes présents sur certaines plateformes et qui regroupent, sous un même identifiant, plus d'un seul contributeur.

Plateformes	**Sources**	**Posts**	**Auteurs**	**% Posts/Source**	**Posts/Auteur**
Forums	110	440 749 318	8 320 125	74.733 %	52,97
Blogs	6	72 315 589	823 871	22.774 %	87,78
Réseaux-Sociaux	2	14 696 161	1 367 415	2.491 %	10,75
Total	120	527 761 068	10 511 411	100,00%	50,21
Corpus de test	115	47 651	2 194	-	21,71

TABLE 1 – Volumétrie des posts par auteurs et par sources

Le corpus de test est un sous ensemble sélectionné de sorte qu'il corresponde au mieux au corpus global (échantillonage d'un rapport d'environ 1 sur 10 000 au regard de l'ensemble du corpus de travail). Le tableau 1 ci-dessus dresse, à titre indicatif, la volumétrie des posts au global et pour le corpus de test.

4 Méthode et applications

Le travail présenté ici ambitionne une couverture à grande échelle et a pour objectif la mise au point d'un modèle robuste de profilage socio-démographique des utilisateurs du web social. Cette problématique de classification nécessite, de part sa complexité, d'aller au delà d'une approche basée sur le repérage de classes fermées de marqueurs linguistiques. En effet, s'il est vrai que certaines expressions employées par les internautes informent -de manière plus ou moins fiable- sur leur sexe ou leur âge, il n'en demeure pas moins qu'une importante partie de signes pertinents se présente de manière diffuse sous forme de signaux faibles, inhérents à l'orthographe ou imbriqués entre les niveaux lexical, morpho-syntaxique et sémantique, ce qui rend ardu leur repérage et laborieuse leur formalisation.

La classification proposée ici s'appuie sur une chaîne de traitements en deux étapes successives :
- Phase 1 à base linguistique :
  - normalisation des posts (débalisage, nettoyage etc.)
  - classification des posts par motifs linguistiques
  - propagation des classes à l'ensemble des posts des utilisateurs qualifiés
- Phase 2 à base d'apprentissage :
  - création d'un modèle bayésien avec les résultats de qualification initiale
  - mise à jour incrémentale du modèle à travers la classification des nouveaux posts / utilisateurs
  - amélioration du modèle par itération sur l'ensemble des données jusqu'à stabilisation

4.1 Contraintes

La tâche de prédiction des profils se confronte à des difficultés techniques liées à la qualité des données, non structurées mais aussi à leur quantité, de plus en plus vaste et évolutive. Le choix de la méthode de classification doit tenir compte de ces paramètres afin d'atteindre un niveau de classification satisfaisant tout en maintenant l'exigence de robustesse dans un contexte applicatif industriel. Un ensemble de facteurs d'ordre textuel, logique et technique sont également à prendre en compte, à savoir :

– qualité des données : textes non structurés, orthographe relâchée, vocabulaires spécifiques
– corrélations entre attributs de certains indicateurs : un célibataire a une probabilité plus forte qu'il soit jeune (avec parfois des cas de contradictions)
– évolution des profils dans le temps : études, mariage, retraite, comptes d'utilisateurs supprimés etc.

Par ailleurs, l'impératif d'avoir à disposition une base d'entraînement supervisée suffisamment large dans notre contexte et compte tenu du coût humain que cela implique, nous avons été amené à en constituer une quasi-automatiquement à base de motifs linguistiques.

4.2 Étape 1 : Classification à base linguistique

Après une phase préalable de filtrage et de normalisation des données, cette étape permet, à l'aide de filtres d'expressions régulières scrupuleusement élaborées, d'identifier la classe à laquelle appartiennent les posts d'un utilisateur donné. Une pondération par score est intégrée à cette classification en vue de différencier les poids des expressions selon leur degré de fiabilité.

4.2.1 Estimation du genre

Les expressions les plus caractéristiques du genre se réfèrent généralement aux attributs sexuels, au statut familial, à la tenue vestimentaire, mais aussi à certaines constructions grammaticales. Voici une liste de quelques exemples d'expressions utilisées, comportant chacune un score de pertinence et déclinées en expressions régulières afin de tenir compte des variations d'écriture et des éventuelles erreurs d'orthographe :

– je suis un(e) : femme/fille/homme/père/mère etc.
– ma femme, mon homme, mon mari, mon époux, mon épouse etc.
– mes seins, mes testicules, mon gynéco, pillule, ivg etc.
– accouchement, césarienne, allaitement etc.
– ma robe, ma jupe, mon sac, mon soutien-gorge etc.
– je + [être] + adj. féminin (par exemple : je suis contente, fâchée, courageuse etc.)
– m' + [avoir] + participe passé (par exemple : il m'a trompée)

4.2.2 Estimation de l'âge

Contrairement au genre qui se présente comme une catégorie binaire discrète, l'âge se traduit par un continuum de valeurs. Cet indicateur peut être exprimé explicitement avec des expressions comme : « j'ai $$ ans », « je suis trentenaire » ou tacitement au travers d'indices socioprofessionnels par exemple : majorité, études, mariage, retraite, etc. Dans ce second cas, des valeurs correspondant aux

tranches d'âges définie (cf. tableau 2) sont retenues. Un certain nombre d'exclusions sont à prendre en compte afin d'éviter des erreurs comme dans l'exemple : « je fume depuis que j'ai 15 ans ».

Classe	Tranche d'âge
Ados	moins de 18 ans
Jeunes	18 - 30 ans
Actifs	30 - 59 ans
Seniors	60 ans et plus

TABLE 2 – Intervalles des tranches d'âge utilisées

4.2.3 Propagation

L'étape de propagation est consécutive à la classification par post. Elle permet d'attribuer à un utilisateur donné la valeur la plus probable estimée à partir de l'ensemble de ses posts. En cas de présence de classes contradictoires, le choix de la classe la plus fréquente est retenu. Le score entre en compte pour la détermination de la classe la plus probable en cas d'égalité.

4.3 Étape 2 : Classification à base d'apprentissage

Cette étape se sert des résultats de la classification initiale en tant que base d'entraînement en vue de construire un modèle de prédiction basé sur un apprentissage incrémental semi-supervisé. Ce modèle reçoit en entrée un post et en sortie lui associe une classe. Ce résultat sert également à la mise à jour de la base de connaissance du modèle de classification.

4.3.1 Choix du modèle bayésien incrémental

Le modèle bayésien multinomial de la librairie Weka (Hall *et al.*, 2009) permet de s'abstraire des éventuelles contraintes liées aux volumes grâce à l'implémentation de l'interface « updatable [4] ». Le test de performance (cf. tableau 3) appliqué à un corpus de 714 posts classés suivant le genre de leurs auteurs, a aidé au choix de ce modèle.

En plus de sa robustesse, le modèle bayésien offre également une lisibilité au niveau des attributs et de la répartition de leurs poids par classe, ce qui aide à la compréhension des résultats de classification. En considérant qu'un entraînement approfondi -permettant la détection des « signaux faibles »- ne devient efficient qu'à partir d'une conséquente volumétrie des données d'apprentissage, ce modèle s'avère bien adapté. De plus, l'apprentissage incrémental a ceci d'avantageux qu'il peut être arrêté et relancé à tout moment tout en restant disponible à la tâche de classification.

4. Possibilité de mettre à jour le modèle incrémentalement lors du processus d'entraînement

Modèle	Tokenizeur	Corrects	précision	Durée [5]	updatable
J48	Token	594	94,14 %	0m :22s	Non
	Ngram :2-2	601	95,24 %	1m :38s	
	Ngram :3-3	395	62,59 %	2m :20s	
	Ngram :1-2	603	95,56 %	2m :18s	
	Ngram :1-3	603	95,56 %	5m :24s	
	Ngram :2-3	601	95,24 %	4m :38s	
NaiveBayes	Token	573	90,81 %	0m :21s	Oui
	Ngram :2-2	568	90,02 %	1m :37s	
	Ngram :3-3	499	79,08 %	2m :23s	
	Ngram :1-2	581	92,07 %	2m :13s	
	Ngram :1-3	582	92,23 %	5m :22s	
	Ngram :2-3	567	89,85 %	4m :40s	
KNN	Token	529	83,83 %	0m :30s	Oui
	Ngram :2-2	546	86,52 %	1m :56s	
	Ngram :3-3	490	77,65 %	2m :38s	
	Ngram :1-2	548	86,84 %	2m :45s	
	Ngram :1-3	537	85,10 %	6m :10s	
	Ngram :2-3	553	87,63 %	5m :09s	
SMO	Token	583	92,39 %	0m :23s	Non
	Ngram :2-2	600	95,08 %	1m :46s	
	Ngram :3-3	513	81,29 %	2m :32s	
	Ngram :1-2	604	95,72 %	2m :17s	
	Ngram :1-3	604	95,72 %	5m :24s	
	Ngram :2-3	593	93,97 %	4m :37s	

TABLE 3 – Comparatif de 4 modèles Weka - Test pour la détection du genre sur 714 posts

4.3.2 Paramètres de la base d'entraînement

Le choix des données d'entraînement résulte de la sélection des posts ayant prioritairement les scores les plus élevés. Les choix suivants ont été comparés afin de choisir le sous-ensemble d'apprentissage le plus pertinent. Le choix 3 a été retenu :

1. données qualifiées : uniquement les posts ayant servi à la qualification
2. données propagées : l'ensemble des posts appartenant à un utilisateur
3. données mixtes : sélection d'une partie des posts qualifiés (10% à 30%) + une partie des posts de propagation.

Les choix suivants ont également été retenus pour la délimitation de la base d'apprentissage :

– longueur des textes : entre 5 et 300 mots
– priorité aux posts qualifiés ayant les scores les plus élevés
– mise à l'écart des posts des supers contributeurs
Les expérimentations pour la création du modèle ont fait l'objet de divers réglages pour la sélection des paramètres les plus adaptés :
– pré-traitements : texte brute / texte désaccentués

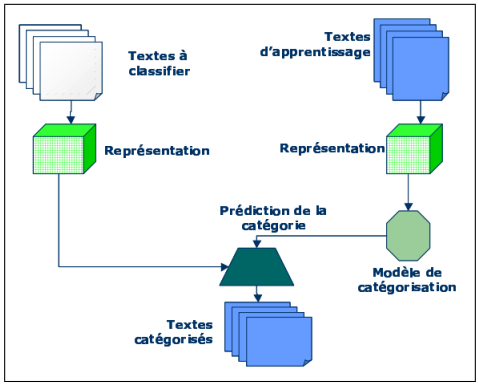

FIGURE 2 – Processus de classification de textes à base d'apprentissage

– tokeniseur : uni-gramme à quadri-grammes
– poids des attributs : binaire / par fréquence
– répartition équitable entre les différentes classes

4.3.3 Sélection d'attributs

La sélection d'attributs permet de conserver un sous ensemble de descripteurs selon leur poids. Le modèle bayésien incrémental que nous utilisons n'offre pas la possibilité de connecter un des algorithmes de sélection d'attributs proposés par Weka. De ce fait, nous avons implémenté notre propre algorithme basé sur un écart-type géométrique (cf. formule ci-après). L'objectif étant de ne conserver que les descripteurs les plus discriminants. Des tests ont été effectué afin de déterminer le seuil le plus pertinent, et ce de manière distincte pour l'âge et pour le genre.

$$\sigma_g = \exp\left(\sqrt{\frac{\sum_{i=1}^{n}(\ln\frac{A_i}{\mu_g})^2}{n}}\right) \qquad (1)$$

5 Résultats et discussions

La classification linguistique a produit les résultats suivants (tableaux 4 et 5) obtenus à l'échelle du corpus global.

Axe	Classe	Posts		Auteurs	
		total	%	total	%
Genre	Masculin	35 102 145	6,65%	2 301 999	21,90%
	Féminin	72 012 369	13,64%	3 899 734	37,10%

TABLE 4 – Taux de couverture de la classification linguistique pour le genre

Axe	Classe	Posts		Auteurs	
		total	%	total	%
Age	Ados	14 960 886	2,83%	511 208	4,86%
	Jeunes	32 736 106	6,20%	1 118 581	10,64%
	Actifs	24 934 811	4,72%	852 013	8,11%
	Seniors	19 402 947	3,61%	650 599	6,19%

TABLE 5 – Taux de couverture de la classification linguistique pour l'âge

Ces résultats ont permis la sélection des corpus de référence pour l'âge et le genre sur la base des scores des utilisateurs les plus élevés. Ainsi, pour un utilisateur donné, cette sélection englobe une partie des posts comportant les motifs linguistiques les plus pertinents ainsi qu'une partie des posts ne comportant aucun de ces motifs. En voici les volumes par posts et par auteurs (cf. tableaux 6 et 7) :

Axe	Classe	Posts		Auteurs	
		total	%	total	%
Genre	Masculin	502 773	19,27%	21 547	30,43%
	Féminin	2 106 614	80,73%	49 256	69,57%
	Total	2 609 387	100,00 %	70 803	100,00 %

TABLE 6 – Volumes des posts « genre » du corpus de référence

Axe	Classe	Posts		Auteurs	
		total	%	total	%
Age	Ados	267 564	20,49%	5 405	17,32%
	Jeunes	567 801	43,49%	11 455	36,71%
	Actifs	374 539	28,69%	8 051	25,80%
	Seniors	95 725	7,33%	6 294	20,17%
	Total	1 305 629	100,00 %	31 204	100,00 %

TABLE 7 – Volumes des posts « âge » du corpus de référence

L'apprentissage bayésien a permis la sélection de 206 314 descripteurs discriminants pour le genre et 153 248 pour l'âge. En voici quelques exemples n'ayant pas forcément les poids les plus élevés (cf. tableaux 8, 9) :

Masculin	Féminin
ferrari	chez son père
ribery	des ballerines
au barca	robe de mariée
taxation	bisous
rendement	henné
chirac	karité
ma calvitie	mon vernis
jeune entrepreneur	une pincée de sel

TABLE 8 – Quelques descripteurs représentatifs du genre

Ados	Jeunes	Actifs	Seniors
dolls	ovulation	gygy	cram
redouble	en bts	hotel	bonne croisière
xd	on est ensemble	bb2	camping cariste
ai les cheveux	fac de	cotisations	avons séjourné
tag	autres filles	congé parental	gendre

TABLE 9 – Quelques descripteurs représentatifs des classes d'âge

5.1 Évaluation

Le gain apporté par la classification à base d'apprentissage par rapport à la méthode linguistique a fait l'objet d'une évaluation sur un échantillon du corpus de test. Les 2 méthodes étant complémentaires, cette mesure permet simplement d'illustrer la marge l'amélioration apporté par l'apprentissage aussi bien pour le genre que pour l'âge. Cette phase étant actuellement en cours d'achèvement, les résultats -encore partiels- sont présentés ici à titre indicatif (cf. tableau 10, 11, 12 et 13).

Méthode	Posts	Rappel	Précision	F-Mesure
Linguistique	5637	3,34%	89,25%	46,29%
Apprentissage		96,68%	83,24%	89,96%
Gain		**93,35%**	**-6,01%**	**43,67%**

TABLE 10 – Gain de classification des posts pour le **genre**

Méthode	Auteurs	Rappel	Précision	F-Mesure
Linguistique	129	68,22%	93,02%	80,62%
Apprentissage		89,15%	72,87%	81,01%
Gain		**20,93%**	**-20,16%**	**0,39%**

TABLE 11 – Gain de classification des auteurs pour le **genre**

Méthode	Posts	Rappel	Précision	F-Mesure
Linguistique	1308	2,29%	94,88%	48,59%
Apprentissage		86,24%	78,13%	82,19%
Gain		**83,94%**	**-16,74%**	**33,60%**

TABLE 12 – Gain de classification des posts pour l'**âge**

5.2 Discussion

Au vu des résultats fournis par la classification linguistique au niveau des posts, la faiblesse du taux de rappel, au profit de la précision, est due à la rigidité volontaire des filtres de classification employés pour les deux indicateurs. Aussi, le taux combiné de F-mesure apparaît plus élevée pour le genre -par rapport à l'âge- du fait de sa binarité et à la richesse des expressions qui lui sont associées. L'étape de propagation aux utilisateurs a permis d'augmenter sensiblement ces taux, permettant ainsi d'obtenir dans des délais raisonnables une base d'apprentissage acceptable qualitativement et satisfaisante en quantité. La classification à base d'apprentissage a permis notamment d'augmenter le taux de rappel pour les deux indicateurs. Le processus itératif prévu en dernière étape sur l'ensemble des données devra améliorer le taux précision. Le taux d'erreur plus élevé lors de ce premier cycle de classification, a tendance à baisser au fur est à mesure des nouvelles classifications, à travers la mise à jour du modèle et l'affinage des poids des descripteurs.

Le nombre élevé des descripteurs retenus par le modèle bayésien, sélectionnés sur la base du différentiel des fréquences entre les classes des deux indicateurs, a mis en évidence des spécificités inhérentes à chacune de ces classes. Les marqueurs les plus saillants (cf. tableaux 6, 7) relèvent de situations socio-professionnelles caractéristiques (emploi, mariage, accouchement, retraite etc.) , de centres d'intérêts (sport, voyage, détente). La méthode a permis également de faire ressurgir des signaux faibles -qu'il convient d'analyser par indicateur et par classe- comme le niveau d'orthographe ou l'emploi de constructions linguistiques spécifiques ("kiffer", "relou", smileys).

D'autre part, le choix du post séparé en tant qu'unité de classification au lieu de regroupements de posts par utilisateur a été consécutif au processus initial de catégorisation à base linguistique. Ce choix, qui peut être discuté, s'est avéré positif puisqu'il permet d'une part de s'affranchir des problèmes de volumétrie notamment pour les données des hyper contributeurs, et d'autre part de procéder à une prédiction plus proche de l'âge étant donné que certains utilisateurs publient des posts à des années différentes, donc à des âges différents. Enfin ce choix permet, d'un point de vue technique, de simplifier le processus de classification et sa mise à jour.

Par ailleurs, un traitement spécifique est prévu sur certaines sources comme Twitter afin de prendre en compte les pseudonymes des utilisateurs pour étayer la classification pour le genre notamment. Cela n'a pas été fait ici pour des raisons de généralisation.

Enfin, ce travail, en cours d'achèvement, a permis de proposer une méthode robuste permettant

Méthode	Auteurs	Rappel	Précision	F-Mesure
Linguistique	44	34,09%	93,18%	63,64%
Apprentissage		86,36%	68,18%	77,27%
Gain		**52,27%**	**-25,00%**	**13,64%**

TABLE 13 – Gain de classification des auteurs pour l'**âge**

d'atteindre dans des délais raisonnables un taux de classification satisfaisant au regard des contraintes de volumétrie et de qualité des données. Le processus itératif, dont la performance n'a pu encore être mesurée, intervient en dernière étape afin de corriger les possibles erreurs des premières qualifications et permet par la même occasion d'affiner la base de connaissance en étendant à l'échelle le volume des données d'apprentissage. Ce travail fait actuellement l'objet d'une généralisation en vue du déploiement.

Remerciements

Un grand merci à Damien Nouvel et à Mathieu Valette ainsi qu'à l'équipe Semantiweb.

Références

ARGAMON S., KOPPEL M., PENNEBAKER J. W. & SCHLER J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, **52**(2), 119–123.

BURGER J. D., HENDERSON J., KIM G. & ZARRELLA G. (2011). Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, p. 1301–1309, Stroudsburg, PA, USA : Association for Computational Linguistics.

CHEN J., HUANG H., TIAN S. & QU Y. (2009). Feature selection for text classification with naïve bayes. *Expert Systems with Applications*, **36**(3), 5432–5435.

EENSOO ET VALETTE E. E. M. (2015). Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité.

ESTIVAL D., GAUSTAD T., PHAM S. B., RADFORD W. & HUTCHINSON B. (2007). Author profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING'07)*, p. 263–272.

HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P. & WITTEN I. H. (2009). The weka data mining software : An update. *SIGKDD Explor. Newsl.*, **11**(1), 10–18.

LIN J. (2007). *Automatic author profiling of online chat logs*. PhD thesis, Monterey, California. Naval Postgraduate School.

MARQUARDT J., FARNADI G., VASUDEVAN G., MOENS M.-F., DAVALOS S., TEREDESAI A. & DE COCK M. (2014). Age and gender identification in social media. In *Proceedings of CLEF 2014 Evaluation Labs*, p. 1129–1136.

MUKHERJEE A. & LIU B. (2010). Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, p. 207–217, Stroudsburg, PA, USA : Association for Computational Linguistics.

NADEAU D. & SEKINE S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, **30**(1), 3–26.

NGUYEN D., SMITH N. A. & ROSÉ C. P. (2011). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, p. 115–123, Stroudsburg, PA, USA : Association for Computational Linguistics.

NGUYEN D., TRIESCHNIGG D., DOGRUÖZ A., GRAVEL R., THEUNE M., MEDER T. & DE JONG F. (2014). *Why Gender and Age Prediction from Tweets is Hard : Lessons from a Crowdsourcing Experiment*, In *Proceedings of COLING 2014, the 25th Conference on Computational Linguistics*, p. 1950–1961. Association for Computational Linguistics (ACL).

OTTERBACHER J. (2010). Inferring gender of movie reviewers : exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, p. 369–378 : ACM.

RANGEL F., ROSSO P., POTTHAST M., STEIN B. & DAELEMANS W. (2015). Overview of the 3rd author profiling task at pan 2015. In *CLEF*.

RANGEL F., STAMATATOS E., MOSHE KOPPEL M., INCHES G. & ROSSO P. (2013). Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, p. 352–365 : CELCT.

RASTIER F. (2009). *Sémantique interprétative*. Presses universitaires de France.

SANTOSH K., BANSAL R., SHEKHAR M. & VARMA V. (2013). Author profiling : Predicting age and gender from blogs—notebook for pan at clef 2013. In *In Forner et*, p.1̃0.

WEREN E. R., KAUER A. U., MIZUSAKI L., MOREIRA V. P., DE OLIVEIRA J. P. M. & WIVES L. K. (2014). Examining multiple features for author profiling. *Journal of Information and Data Management*, **5**(3), 266.