Analyse de l'image de marque sur le Web 2.0

Jean-Valère Cossu

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 536 « *Sciences et Agrosciences* »
Laboratoire d'Informatique (EA 931)

Analyse de l'image de marque sur le Web 2.0

par

Jean-Valère COSSU

Soutenue publiquement le 16 décembre 2015 devant un jury composé de :

M^{me}	Josiane Mothe	Professeur, IRIT Toulouse	Rapporteur
M.	Philippe Mulhem	Chargé de Recherche CNRS (HDR), LIG Grenoble	Rapporteur
M.	Jian-Yun Nie	Professeur, Université de Montréal, Canada	Examinateur
M.	Frédéric Gimello-Mesplomb	Professeur, Université d'Avignon	Examinateur
M^{me}	Rocio Abascal Mena	Maître de Conférences, UAM Mexico	Examinateur
M.	Paolo Rosso	Associate Professor, Politècnica Valencia	Examinateur
M.	Julien Velcin	Maître de Conférences (HDR), Université Lyon II	Examinateur
M.	Marc El-Bèze	Professeur, Université d'Avignon	Directeur de thèse

Ce travail a également été encadré par :

M.	Juan-Manuel Torres	Maître de Conférences (HDR), Université d'Avignon	Co-Directeur de thèse
M.	Eric SanJuan	Maître de Conférences, Université d'Avignon	Co-encadrant

Membres invités :

M.	Jussi Karlgren	Senior Researcher, Gavagai, Stockholm	Invité
M.	Vincent Labatut	Maître de Conférences, Université d'Avignon	Invité
M.	Christophe Navas	Directeur R&D, Vodkaster	Invité

Thèse réalisée dans le cadre du projet ANR Imagiweb avec le soutien de Vodkaster

---

Laboratoire Informatique d'Avignon

Remerciements

Je tiens à remercier l'ensemble des membres de mon jury, pour avoir accepter de partager leurs regards attentifs sur mon travail, la qualité de nos échanges et bien sûr d'avoir fait le déplacement jusqu'en Avignon.

Je voudrais ensuite chaleureusement remercier mes encadrants, à commencer par Marc El-Bèze pour m'avoir donné ma chance lors cette alternance durant de ma dernière année de Master, pour l'aide à la rédaction et les conseils distillés au cours de ces quatre années passées ensemble dont les derniers mois à partager le même bureau. Je remercie ensuite Juan-Manuel Torres qui m'avait prévenu en début de thèse, au bout de trois ans je finirai par parler espagnol. Nous n'en sommes pas encore la mais presque ! Je remercie également Eric San-Juan pour son ouverture d'esprit, son esprit critique et cette autre vision de la recherche qu'il a su me montrer ainsi que son analyse de mon travail.

Je souhaite également remercier les membres du projet Imagiweb pour ces regards croisés sur nos différents domaines de recherche ainsi que mes collègues de travail au LIA pour la bonne ambiance de travail et nos discussions parfois longues et perchées. Merci aussi à mes nouveaux collègues de Vodkaster avec qui ça déménage au quotidien.

Plus personnellement, je tiens à exprimer mes pensées à mes camarades de l'alliance francophone pour toutes ses soirées endiablées à refaire le monde. Merci également François pour ton amitié, ton soutien, et tous ces excellents moments passés en ta compagnie.

Enfin, je remercie ma famille, plus particulièrement mes parents pour l'amour et la confiance dont ils m'ont toujours témoigné, c'est une fierté pour moi d'en arriver la aujourd'hui. Bien sûr, mes remerciements ne seraient pas complets si je n'insistais pas sur le grand mérite de Marie ma délicieuse épouse, pour avoir partagé ma vie de doctorant pleine de doutes mais aussi de satisfactions. Merci d'avoir eu la patience de relire la thèse mais aussi et pour cette aventure en Avignon, sans toi toutes ces années n'auraient pas été les mêmes.

Une page de ma vie qui se tourne pour une nouvelle aventure qui commence.

Résumé

En plus d'être un moyen d'accès à la connaissance, Internet est devenu en quelques années un lieu privilégié pour l'apparition et la diffusion d'opinions. Chaque jour, des millions d'individus publient leurs avis sur le Web 2.0 (réseaux sociaux, blogs, etc.). Ces commentaires portent sur des sujets aussi variés que l'actualité, la politique, les résultats sportifs, biens culturels, des objets de consommation, etc. L'amoncellement et l'agglomération de ces avis publiés sur une entité (qu'il s'agisse d'un produit, une entreprise ou une personnalité publique) donnent naissance à l'image de marque de cette entité.

L'image d'une entité est ici comprise comme l'idée qu'une personne ou qu'un groupe de personnes se fait de cette entité. Cette idée porte a priori sur un sujet particulier et n'est valable que dans un contexte, à un instant donné. Cette image perçue est par nature différente de celle que l'entité souhaitait initialement diffuser (par exemple via une campagne de communication). De plus, dans la réalité, il existe au final plusieurs images qui cohabitent en parallèle sur le réseau, chacune propre à une communauté et toutes évoluant différemment au fil du temps (imaginons comment serait perçu dans chaque camp le rapprochement de deux hommes politiques de bords opposés). Enfin, en plus des polémiques volontairement provoquées par le comportement de certaines entités en vue d'attirer l'attention sur elles (pensons aux tenues ou déclarations choquantes), il arrive également que la diffusion d'une image dépasse le cadre qui la régissait et même parfois se retourne contre l'entité (par exemple, « *le mariage pour tous* » devenu « *la manif pour tous* »). Les opinions exprimées constituent alors autant d'indices permettant de comprendre la logique de construction et d'évolution de ces images. Ce travail d'analyse est jusqu'à présent confié à des spécialistes de l'e-communication qui monnaient leur subjectivité. Ces derniers ne peuvent considérer qu'un volume restreint d'information et ne sont que rarement d'accord entre eux.

Dans cette thèse, nous proposons d'utiliser différentes méthodes automatiques, statistiques, supervisées et d'une faible complexité permettant d'analyser et représenter l'image de marque d'entité à partir de contenus textuels les mentionnant. Plus spécifiquement, nous cherchons à identifier les contenus (ainsi que leurs auteurs) qui sont les plus préjudiciables à l'image de marque d'une entité. Nous introduisons un processus d'optimisation automatique de ces méthodes automatiques permettant d'enrichir les données en utilisant un retour de pertinence simulé (sans qu'aucune action de la part de l'entité concernée ne soit nécessaire). Nous comparer également plusieurs approches de contextualisation de messages courts à partir de méthodes de recherche d'information et de résumé automatique. Nous tirons également parti d'algorithmes de modélisation (tels que la Régression des moindres carrés partiels), dans le cadre d'une modélisation conceptuelle de l'image de marque, pour améliorer nos systèmes automatiques de catégorisation de documents textuels. Ces méthodes de modélisation et notamment les représentations des corrélations entre les différents concepts que nous manipulons nous permettent de représenter d'une part, le contexte thématique d'une requête de l'entité et d'autre, le contexte général de son image de marque. Nous expérimentons l'utilisation et la combinaison de différentes sources d'information générales représentant les grands types d'information auxquels nous sommes confrontés sur internet : de long les contenus objectifs rédigés à des informatives, les contenus brefs générés par les utilisateurs visant à partager des opinions. Nous évaluons nos approches en utilisant deux collections de données, la première est celle constituée dans le cadre du projet Imagiweb, la seconde est la collection de référence sur le sujet : CLEF RepLab.

Mots-clés : Traitement Automatique de la Langue Naturelle Écrite, Recherche d'Information, Contextualisation, Concepts Implicites, Modélisation Thématique, Apprentissage Automatique, Aide à la Décision, Informatique Décisionnelle, E-Reputation, RepLab, Imagiweb.

Abstract

Every day, millions of people publish their views on Web 2.0 (social networks, blogs, etc.). These comments focus on subjects as diverse as news, politics, sports scores, consumer objects, etc. The accumulation and agglomeration of these notices on an entity (be it a product, a company or a public entity) give birth to the brand image of that entity. Internet has become in recent years a privileged place for the emergence and dissemination of opinions and putting Web 2.0 at the head of observatories of opinions. The latter being a means of accessing the knowledge of the opinion of the world population.

The image is here understood as the idea that a person or a group of people is that entity. This idea carries a priori on a particular subject and is only valid in context for a given time. This perceived image is different from the entity initially wanted to broadcast (eg via a communication campaign). Moreover, in reality, there are several images in the end living together in parallel on the network, each specific to a community and all evolve differently over time (imagine how would be perceived in each camp together two politicians edges opposite). Finally, in addition to the controversy caused by the voluntary behavior of some entities to attract attention (think of the declarations required or shocking). It also happens that the dissemination of an image beyond the framework that governed the and sometimes turns against the entity (for example, « *marriage for all* » became « *the demonstration for all* »). The views expressed then are so many clues to understand the logic of construction and evolution of these images. The aim is to be able to know what we are talking about and how we talk with filigree opportunity to know who is speaking.

In this thesis we propose to use several simple supervised statistical automatic methods to monitor entity's online reputation based on textual contents mentioning it. More precisely we look the most important contents and theirs authors (from a reputation manager point-of-view). We introduce an optimization process allowing us to enrich the data using a simulated relevance feedback (without any human involvement). We also compare content contextualization method using information retrieval and automatic summarization methods. We also propose a reflection and a new approach to model online reputation, improve and evaluate reputation monitoring methods using Partial Least Squares Path Modelling (PLS-PM). In designing the system, we wanted to address local and global context of the reputation. That is to say the features can explain the decision and the correlation betweens topics and reputation. The goal of our work was to propose a different way to combine usual methods and features that may render reputation monitoring systems more accurate than the existing ones. We evaluate and compare our systems using state of the art frameworks : Imagiweb and RepLab. The performances of our proposals are comparable to the state of the art. In addition, the fact that we provide reputation models make our methods even more attractive for reputation manager or scientists from various fields.

Keywords: Natural Language Processing, Information Retrieval, Contextualization, Implicit Concepts, Modelling, Machine Learning, Artificial Intelligence, Business Intelligence, E-Reputation, RepLab, Imagiweb.

Publications de l'auteur

Revues internationales

1. ***(24p) A Review of Features for the Discrimination of Twitter Users : Application to the Prediction of Offline Influence***. *Cossu J-V., Labatut V. and Dugue N.* : SNAM Special Issue on Diffusion of Information and Influence in Social Networks (2016).
2. *(8p) Bilingual and Cross Domain Politics Analysis*. *Cossu J.-V, Abascal R., Molina A., Torres-Moreno J. M. and SanJuan E.* : Research in Computing Science (ISSN 1870-4069) Issue 85 (2014), page 9-19.

Conférences et Workshops internationaux

1. ***(Long 11p) Intweetive Text Summarization***. *Cossu J-V., Torres-Moreno, J. M, San-Juan E.and El-Bèze M.* : 14th Mexican International Conference on Artificial Intelligence (MICAI), (Mexico (Mexique) Octobre 25-31 2015).
2. *(Court 6p) Multi-Dimensional Reputation Modeling using Micro Blog contents*. *Cossu J-V., San-Juan E., Torres-Moreno, J. M and El-Bèze M.* : 22nd International Symposium on Methodologies for Intelligent Systems (ISMIS), (Lyon (France) Octobre 21-23 2015).
3. ***(Long 8p) Detecting Real-World Influence Through Twitter***. *Cossu J-V., Dugue N. and Labatut V.* : The Second European Network Intelligence Conference (ENIC), (Karlskrona (Suède) Septembre 21-22 2015).
4. ***(Long 12p) NLP-based classifiers to generalize experts assessments in E-Reputation***. *Cossu J-V., Ferreira E., Gaillard J., Janod K. and El-Bèze M.* : Sixth International Conference of the CLEF initiative (Toulouse (France) Septembre 8-11 2015).

5. *(Long 10p) Machine Learned Annotation of tweets about politicians' reputation during Presidential Elections : the cases of Mexico and France. Cossu J.-V, Abascal R., Molina A., Torres-Moreno, J. M. and San-Juan, E.* : Workshop on Replicability and Reproducibility in Natural Language Processing : Adaptive methods, resources and software at IJCAI 2015,(Buenos Aires (Argentine) Juillet 25-27 2015).
6. ***(Court 6p) Automatic Classification and PLS-PM Modeling for Profiling Reputation of Corporate Entities on Twitter**. Cossu J-V., San-Juan E., Torres-Moreno, J. M and El-Bèze M.* : 20th International Conference on Application of Natural Language to Information Systems (NLDB), (Passau (Allemagne) Juin 17-19 2015).
7. *(Long 10p) An opinion mining Partial Least Square Path Modeling for football betting. El Hamdaoui M. and Cossu J-V.* : PhD Session of the 7th European Conference on Machine Learning and Practice of Knowledge Discovery in Databases (Nancy (France) Septembre 15-19 2014).
8. ***(Court 6p) Towards the improvement of topic priority assignment using various topic detection methods for e-reputation monitoring on Twitter**. Cossu J-V., Bigot B., Bonnefoy L. and Senay G.* : 19th International Conference on Application of Natural Language to Information Systems (NLDB), (Montpellier (France) Juin 18-20 2014).

Conférences et Workshops nationaux

1. *(Demo 2p) Etude de l'image de marque d'entités dans le cadre d'une plateforme de veille sur le Web social. Khouas L., Brun C., Peradotto A., Cossu J-V., Boyadjian J. and Velcin J.* : 22ème Conférence sur le Traitement Automatique des Langues Naturelles, (DEFT/TALN 2015), (Caen (France) June 22-25 2015).
2. *(Court 8p) Recherche et utilisation d'entités nommées conceptuelles dans une tâche de catégorisation. Cossu J-V., Torres-Moreno J-M. and El-Bèze M.* : 20ème Conférence sur le Traitement Automatique des Langues Naturelles, (DEFT/TALN 2013), (Sables d'Olonne (France) Juin 17-21 2013).
3. *(Long 10p) Contextualisation de messages courts : l'importance des métadonnées. Cossu J-V., Gaillard J., Torres-Moreno J-M. and El-Bèze M.* : 13ème Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2013), (Toulouse (France) Janvier 28 2013).

Publications et rapports techniques

1. *A Review of Features for the Discrimination of Twitter Users : Application to the Prediction of Offline Influence*. *Cossu J-V., Labatut V. and Dugue N.* : arXiv preprint arXiv :1509.06585 (2015).
2. *How to merge three different methods for information filtering*. *Cossu J-V., Bonnefoy L., Bost X. and El-bèze M.* : arXiv preprint arXiv :1510.07385 (2015).
3. *An Author-Topic based Approach to Cluster Tweets and Mine their Location*. *Morchid M., Portilla Y., Josselin D., Dufour R., Altman E., El-Beze M., Cossu J-V., Linarès G., and Reiffers-Masson A.* : Procedia Environmental Sciences, 27, 26-29. (2015).
4. *Analyser l'image de marque d'entités sur le web. Revue du projet ImagiWeb.*. *Velcin J., Peradotto A., Khouas L., Cossu J-V., Dormagen J-Y. and Brun C.* : Ingénierie des Systèmes d'Information 19(3) : 159-162 (2014).
5. *LIA@RepLab 2014*. *Cossu J-V., Ferreira E., Gaillard J., Janod K. and El-Bèze M.* : Working Notes for Fifth International Conference of the CLEF initiative (Sheffield (UK) Septembre 15-18 2014).
6. *E-reputation monitoring on Twitter with active learning automatic annotation*. *Cossu J-V., El-Bèze M., Torres-Moreno, J. M and San-Juan E.* <hal-01002818> (2014).
7. *LIA@RepLab 2013*. *Cossu J-V., Bigot B., Bonnefoy L., Morchid M., Bost X., Senay G., Dufour R., Bouvier V., Torres-Moreno J-M and El-bèze M.* : Working Notes for Fourth International Conference of the CLEF initiative (Valencia (Espagne) Septembre 23-26 2013).
8. *Systèmes du LIA à DEFT'13*. *Bost X., Brunetti I., Cabrera-Diego L-A., Cossu J-V., Linhares A., Morchid M., Torres-Moreno J-M., El-bèze M. and Dufour R.* : Actes du neuvième Dèfi Fouille de Textes (DEFT/TALN), (Sables d'Olonne (France) Juin 17-21 2013).

Table des matières

Liste des illustrations

Liste des tableaux

Chapitre 1

Introduction

Sommaire

Introduction

Nous nous intéressons dans cette thèse à l'analyse des *images de marque d'entités de toute sorte sur le Web 2.0*, également appelée *E-réputation*. L'encyclopédie en ligne *Wikipédia* définit l'image de marque comme «*la représentation perçue par le public d'une personnalité, d'une entreprise, d'une organisation, d'une institution, de leurs produits et de leurs marques commerciales*». Cette notion trouve ses racines dans le contexte des premiers échanges commerciaux, époque où l'on commença à miser sur l'hypothèse qu'un produit serait préféré à un autre car il avait meilleure *réputation*. L'image de marque est issue des multiples relations entre plusieurs entités et de la combinaison de leurs représentations individuelles. Ces représentations sont immatérielles et difficilement quantifiables, et même si elles peuvent revêtir plusieurs formes matérielles : bouche-à-oreille, discours construit écrit ou oral, différents médias de communications (supports publicitaires), chacun reste libre (quoi que sous influence) dans la perception qu'il a de cette représentation. Plusieurs facteurs sont donc à prendre en compte lorsque l'on souhaite analyser l'image de marque d'une entité :

1. l'émetteur de cette image et ses intentions : qui communique ? Dans quel but ?
2. le contenu de l'image : quelles sont les thématiques abordées ? Sous quel angle le sont-elles ?
3. les récepteurs de l'image, leurs perceptions : à qui est destinée la communication ? Le message transmis est-il compris ?

Toutefois, il n'existe pas vraiment de consensus sur une réelle définition des éléments qui composent l'image de marque, à l'instar des critères qui permettraient justement de comparer les images respectives de deux entités concurrentes. De plus, les moyens de télécommunications modernes ont profondément modifié la manière dont nous consommons l'information. Les médias sociaux tels que Twitter [1], Facebook [2] et autres services de «*blogging*» [3] sont utilisés par des millions de personnes, qui créent et partagent librement du contenu. Dans notre cas, ce contenu correspond plus précisément à de l'information liée à divers types d'événements, que ces derniers soient privés ou public et, d'importance variable.

1. https://twitter.com/
2. https://facebook.com/
3. Sont considérés comme tels tous les services permettant à leurs utilisateurs de partager des contenus qu'ils eux mêmes générés et ce quelque soit leurs formes.

Le récent développement des médias sociaux ainsi que leur capacité à influencer la société amènent l'analyse d'opinions (souvent également qualifiée d'analyse de sentiments) à faire l'objet d'un engouement tout particulier dans certains milieux comme la recherche académique ou l'industrie. En effet, l'émergence de ces plate-formes de discussions compilant des avis utilisateurs permet d'accéder à une masse très importante de documents contenant des informations exprimant des opinions. Ces collections constituent de fait une source énorme de données qui peut être exploitée par toutes les applications de veille (technologique, marketing, culturelle, politique concurrentielle, sociétale) (Aggarwal, 2011) pour accéder aux opinions exprimées spontanément par les internautes sur différentes entités comme des personnalités, des entreprises ou des marques de produits. L'analyse de ces données permet d'identifier l'image de l'entité telle qu'elle est perçue ainsi que son positionnement par rapport à d'autres entités. Cette connaissance est précieuse pour mener les actions adaptées en vue de maintenir une bonne image, corriger certains aspects et prévenir d'éventuelles crises d'image permettant la mise en place d'un cycle de veille complet dans des contextes divers tels que l'intelligence économique et l'e-réputation. Ce cycle de veille peut se décomposer en quatre grandes phases :

1. l'acquisition de l'information grâce à des mécanismes de collecte automatique des données,
2. le traitement de ces données collectées et leur intégration dans une base de données d'entreprise
3. l'extraction de l'information pertinente,
4. le partage et la diffusion de l'information auprès des collaborateurs concernés par la gestion de l'image de marque de l'entreprise ?

Dans le cadre du projet Imagiweb, l'analyse de l'image de marque se concentre sur le réseau social Twitter. A ce titre, nous laisserons de coté la première phase pour nous concentrer sur les trois autres. En effet, la collecte d'information sur les médias sociaux et plus particulièrement sur Twitter fait l'objet d'interrogations quant à la réelle qualité des contenus auxquels nous avons accès (Gerlitz et Rieder, 2013) ainsi que la représentativité des utilisateurs (Boyadjian, 2014).

L'informatique, en tant que science du traitement automatisé des informations cherche à apporter des solutions aux problèmes que pose le suivi de ces images. Il existe dès lors un nombre important d'outils informatiques proposant de suivre et analyser le flux d'informations lié à une marque. Ces outils sont parfois proposés par les plateformes sociales, comme le fait par exemple Twitter avec son service d'analyse des tendances.

Néanmoins, le problème aujourd'hui n'est plus la disponibilité de l'information, mais plutôt la disponibilité de notre attention à nous, les potentiels récepteurs (Simon, 1971). Car si toutes ces connaissances peuvent être plus ou moins intégrées par des systèmes automatiques de traitement de l'information, il est humainement impossible de pouvoir les appréhender dans leur globalité. Devant l'abondance des données disponibles [4], il devient alors nécessaire si ce n'est impératif de développer des outils permettant une exploitation efficace et rationnelle de l'information circulant dans ces médias sociaux.

Outre les nombreux travaux issus de la sociologie, la psychologie cognitive et des sciences de la communication et de l'information, plusieurs domaines de l'informatique s'intéressent également à ces problèmes. Citons particulièrement les domaines suivants :

— la fouille de données qui consiste à collecter des données, les pré-traiter puis les transformer dans le but de les soumettre à des algorithmes pour les analyser ;
— l'ingénierie des connaissances qui pourrait correspondre aux outils appliqués sur le résultat (connaissances extraites) des algorithmes cités précédemment ;
— sur un type de données bien plus spécifique, citons également les méthodes de traitement automatique de la langue naturelle et de recherche d'information.

L'objectif de notre travail est d'analyser les volumes de données générés par les médias sociaux et de développer des outils afin de mieux comprendre la dynamique de ces images et ainsi améliorer notre rapport à l'information (Aggarwal, 2011). Des travaux suivant une orientation plus sociologique ou centrée sur les échanges (approches de types «*réseaux*») s'intéressent à la catégorisation des utilisateurs (comprendre les rôles émetteurs et récepteurs de l'image) en analysant par exemple la structure du réseau d'interconnexion entre les différents groupes d'utilisateurs. Déterminer des groupes d'utilisateurs similaires connaissant ces différentes communautés permet de leur proposer des éléments d'informations plus pertinents susceptibles de les intéresser davantage.

4. L'information touche d'ailleurs un nombre toujours croissant de personnes (Bakshy et al., 2012).

Processus de traitement

La figure 1.1 montre les principales étapes généralement considérées dans les processus d'analyse d'image de marques, pour chaque étape nous précisons quelques unes des méthodes les plus utilisées. Dans le cadre du projet Imagiweb, ces étapes sont la collecte d'informations, la création des données, le pré-traitement des données, leur analyse, l'interprétation des résultats et les éventuelles réactions.

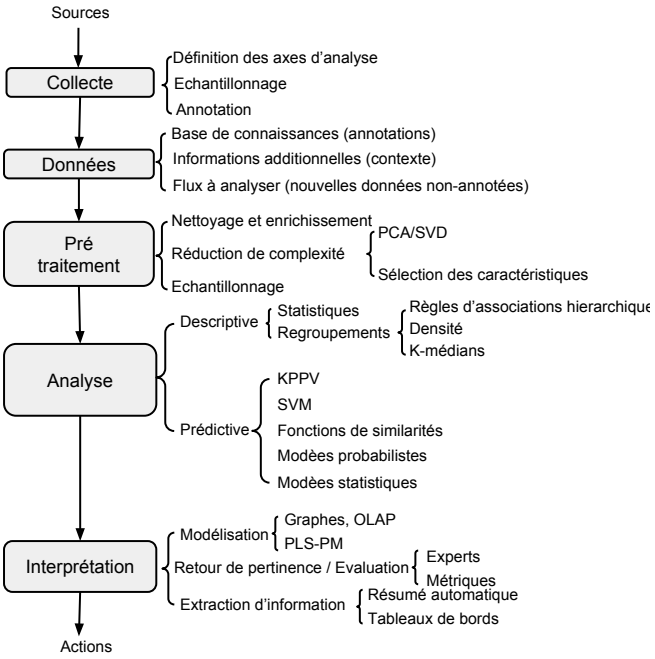


FIGURE *1.1* – *Processus d'analyse de l'image de marque dans la cadre du projet Imagiweb.*

1.1 Problématiques et contributions

Les contributions apportées par nos travaux dans le cadre de cette thèse se concentrent sur trois problématiques qui découlent du phénomène de diffusion de l'information dans les médias sociaux. Nous nous intéressons plus particulièrement à l'analyse de l'information véhiculant l'image de marque au travers de trois problématiques :

1. analyser le contenu des messages afin de détecter les éléments d'information qui suscitent intérêt et réaction des utilisateurs,
2. identifier les différentes catégories d'utilisateurs dont ceux ayant une influence sur la diffusion et la perception de l'information,

3. modéliser l'image de marque et présenter la manière dont elle est perçue à un analyste afin de l'assister.

Au travers de ces problématiques, notre objectif réside dans le développement d'applicatifs permettant l'annotation automatique des collections de données. Cela consiste à identifier des opinions et à en extraire les différentes caractéristiques conformément aux modèles de connaissances représentant l'image de marque d'une entité. Ces contributions, que nous positionnons dans la littérature et que nous décrivons brièvement dans la suite de cette section, sont formulées de manière générique par rapport aux médias sociaux, et évaluées par rapport à un média social en particulier, à savoir Twitter [5]. Même si ces notions sont maintenant relativement communes, avant d'aller plus loin nous proposons de rappeler quelques concepts propres à *Twitter*. Plus de détails ainsi que des informations complémentaires sur d'autres services de médias sociaux sont disponibles dans les travaux de (Ellison et al., 2007).

Twittere est un service de « *micro-blogging* » en ligne qui permet à ses utilisateurs d'échanger en temps réel autour de n'importe quel sujet à partir de messages appelés « *tweets* » (des messages dont la taille ne dépasse pas 140 caractères). Chaque tweet publié apparaît sur la page de profil de son auteur. Pour voir l'ensemble des messages d'un utilisateur il suffit donc de se rendre sur la page de profil de ce denier. De même, afin de faciliter l'accès aux messages des autres utilisateurs, les utilisateurs ont la possibilité de s'abonner les uns aux autres (on devient alors un « *follower* », c'est l'action de suivre un utilisateur, l'utilisateur suivi devient alors le « *followee* »). Dans ce cas, chaque message publié par un utilisateur et est instantanément transmis à ses abonnés, ces derniers peuvent avoir une vue des l'activité de leurs « *abonnements* » dans leur « *timeline* ». Celle-ci consiste en l'empilement en ordre chronologique inverse des tweets publiés par les utilisateurs auxquels nous sommes abonnés. Autre particularité intéressante, il est possible de re-diffuser le message d'un autre utilisateur via l'action de « *retweet* » dans le but de le partager avec ses abonnés pour mentionner son accord (ou son désaccord si l'on rajoute un commentaire) avec le message. Dernière caractéristique, les contenus peuvent comporter deux symboles particuliers en préfixe de certains mots. Pour mentionner spécifiquement un utilisateur, on utilise « *@NomUtilisateur* » [6]. Il est possible de lancer des discussions autour d'un mot (ou d'une expression) qui devient alors un « *mot-dièse* » (ou « *hashtag* »). Ces hashtags qui sont une chaîne de caractères commençant par le caractère dièse « *#motouexpression* » ont fait l'objet de nombreuses études (Brun et Roux, 2014).

5. Une évaluation additionnelle à partir de billets de blogs sera également présentée
6. La littérature (Li et al., 2011) indique également cette information de l'utilisateur qui écrit ou à qui est destiné le message est indice important permettant de découvrir l'opinion véhiculée dans ce message.

1.1.1 Analyser les contenus

Problématiques

Etant donné le volume de publications quotidiennes sur les médias sociaux, l'analyse des messages est devenu une tâche de Fouille de Données (FD) (données que l'on peut considérer comme massive). Afin d'être en mesure de réagir rapidement à des publications dangereuse pour notre réputation il st nécessaire de faire appel à des méthodes automatiques issues de plusieurs domaines de recherche comme la Recherche d'Information (RI), l'Extraction et la Gestion de Connaissances (EGC), le Traitement de la Langue Naturelle (TALN), et l'Apprentissage Automatique (AA). Dans le cadre de l'analyse de données issues des médias sociaux, ces différents domaines de recherche partagent entre autres un objectif commun : automatiser au maximum la détection et l'analyse des sujets (ou thématiques généralement appelés «*topics*» sur les médias sociaux (Makkonen et al., 2004)). L'objectif de cette automatisation est de permettre à l'analyse de réputation de se focaliser sur la prise de décision (faut-il réagir suite à cette publication ?).

Chaque domaine de recherche définit des méthodes qui, de part des analyses statistiques et sémantiques, permettent d'apprendre à traiter des contenus textuels. Ces traitements sont effectués à partir de corpus de données textuelles en se basant sur la présence (pondérée ou non) et les (co-)occurrences de certains mots voire d'expressions (suites de mots que l'on appelle communément *n*-grams) ou plus spécifiquement d'«*Entités Nommées*». Il existe bien sûr un grand nombre d'applications à ces analyses de contenus (Kontostathis et al., 2010). Nous pouvons trouver dans la littérature plusieurs types d'approches, celles qui se basent d'une part sur des méthodes d'apprentissage automatique, visant à détecter une caractéristique spécifique, celles dont l'objectif est d'apprendre et reproduire un jugement d'expert, et enfin celles qui reposent sur l'utilisation de connaissances comme les dictionnaires (Miller, 1995) et autres répertoires d'«*Entités Nommées*» (Derczynski et al., 2015). Les performances de ces méthodes sont par contre sensibles quant à la disponibilité et l'évolution des ressources utilisées.

Contributions

Nous proposons plusieurs méthodes statistiques de traitement de la langue par apprentissage automatique à partir d'un jeu réduit de connaissances fournies par des experts pour détecter automatiquement l'opinion (polarité et thématique), ou la priorité de l'information, véhiculée dans les messages à partir du contenu textuel de ces derniers. Ces méthodes relativement simples nous

permettent également de pouvoir facilement extraire l'information qui a contribué à la prise d'une décision afin de permette à l'utilisateur de notre système de pouvoir d'une part comprendre la décision et si besoin est d'effectuer des modifications qui enrichiront le modèle. Les expérimentations menées montrent la pertinence des méthodes proposées via l'évaluation et la comparaison des performances de ces méthodes avec celles qui sont qualifiées «*état de l'art*». Cette évaluation nous place le plus souvent au niveau de ces dernières malgré la complexité bien moindre de nos méthodes.

1.1.2 Détecter les différentes populations d'utilisateurs

Problématiques

Tous les acteurs de la société (entreprises, personnalités plus ou moins connues et individus lambdas) se retrouvent à composer la très grande population d'utilisateurs des médias sociaux. Toutefois, un nombre plus restreint d'entités cherche à analyser la réaction de cette masse afin de promouvoir un nouveau produit, analyser les réactions sur une idée politique, identifier certains types d'utilisateurs comme les opposants politiques, spammeurs, criminels, leaders d'opinion etc.. Historiquement, ces analyses, même si parfois menées plus ou moins officiellement par des services gouvernementaux, sont confiées à un nombre limité d'agences de communications ou de sondages qui se chargent d'aller directement interroger les utilisateurs pour collecter les informations nécessaires. Une fois ces éléments collectés, la principale démarche d'analyse mise en oeuvre par la suite vise à séparer les utilisateurs en différents groupes afin de procéder à une analyse plus fine et ainsi prendre des décisions voire éventuellement agir (dans le cas des criminels) sur une échelle réduite. Pour ce faire, il faut définir précisément ce qu'est un profil utilisateur. La littérature diverge à ce sujet, chaque domaine et groupe de recherche ayant sa définition et utilisant chacune des caractéristiques différentes (avec souvent des problèmes de dénominations, plusieurs équipes associant une même dénomination à des caractéristiques pourtant différentes et «*vice-versa*» une même caractéristique se retrouvant qualifiée sous plusieurs formes) et bien entendu trop rares sont les cas où données et implémentations sont mises à disposition de la communauté. C'est pourquoi nous avons été amenés à réfléchir d'une part sur le partage et la réutilisation de toutes ces connaissances et d'autre part sur la proposition de méthodes automatiques permettant de mener cette analyse.

Contributions

Nous proposons de revenir sur l'ensemble des éléments qu'il est possible de considérer pour caractériser un profil d'utilisateur et de catégoriser ces éléments. Nous proposons également de ramener le profil d'un individu à ce que ce dernier écrit au travers de deux variantes de modélisation d'un profil. Les résultats de cette analyse peuvent également permettre d'améliorer des approches d'analyses automatiques des contenus en plaçant dans le contexte du profil la décision prise au niveau d'un contenu individuel lié à ce profil. Nous évaluons la pertinence de l'ensemble de nos propositions et les comparons à l'état de l'art. A ce titre, nos expériences ont montré que nos méthodes se révèlent être particulièrement efficaces pour répondre à ces problématiques et notamment la détection d'utilisateurs qui semblent être les plus influents ou encore la catégorisation de ces utilisateurs en fonctions de divers critères dont les fameuses « *catégories socioprofessionnelles* » (CSP).

1.1.3 Organisation et présentation de l'information pertinente

Problématiques

Nous avons vu que les utilisateurs des médias sociaux partagent, discutent et retransmettent en temps réel de l'information à propos d'événements divers. De nombreuses études ont déjà montré qu'il était possible d'exploiter les médias sociaux pour détecter automatiquement les événements importants, les analyser et en extraire les éléments qui étaient les plus à même de susciter l'intérêt des utilisateurs. Ces études ont permis de révolutionner la notion de « *veille d'information* » et les métiers associés au journalisme, ainsi qu'au marketing, plus habitués à travailler à partir des médias traditionnels. Aujourd'hui, ces médias comme les services de presse ne peuvent que difficilement rivaliser avec ce volume sans cesse croissant de messages publiés sur les médias sociaux au point d'en arriver aux mêmes pratiques. Ce volume engendre toutefois une surcharge informationnelle. Il devient alors très difficile d'identifier des éléments d'information pertinents liés à des événements spécifiquement importants aux yeux d'une entité, ces éléments étant noyés dans un grand volume de messages sans rapport (« *le bruit* »). Nous ne nous intéresserons pas aux problèmes liés à la vérification des données. En effet, l'immédiateté de l'information sur les médias sociaux incite certains contributeurs à ne pas se soucier de la véracité de l'information qu'ils transmettent.

Au-delà de la détection, *a posteriori*, des détails ayant suscité intérêt et réactions des utilisateurs d'un média social, il est important de déterminer quels sont les éléments d'informations diffusés par les utilisateurs qualifiés d'influents. Il apparaît parfois également utile de pouvoir éviter la diffusion de l'information à une plus grande masse d'utilisateurs et ainsi anticiper leur réaction. Il faut déterminer pourquoi on se doit de réagir à un événement spécifique (un « *tweet* » sur un aspect particulier d'un produit ou service) mais aussi et surtout comment réagir. C'est la communication de crise, l'occupation principale des gestionnaires d'images ou de communautés (« *community manager* »). Cela nous amène à nous demander comment permettre à l'expert (ainsi qu'à des non-experts) d'analyser efficacement l'image d'une marque à partir des données collectées sur les médias sociaux ? Les réponses à ces questions bénéficieraient autant aux non-experts ayant besoin d'analyser les données dont ils disposent, qu'aux chercheurs désirant pousser plus loin une analyse de contenus ou d'image de marque.

Contributions

Modéliser la réputation à travers toutes les interactions imbriquées (opinions, utilisateurs, thèmes tels que définit par (Makkonen et al., 2004)) est une tâche qui reste encore trop peu abordée dans la littérature. Notre contribution est ici double, nous proposons d'utiliser les méthodes statistiques et probabilistes qui seront décrites dans les chapitres que nous venons d'évoquer ci-dessus pour détecter automatiquement les contenus importants[7] qui devraient susciter l'intérêt des entités et analystes soucieux de leur image, à partir des flux de messages publiés dans les médias sociaux en se basant sur le contenu textuel des messages.

Nous proposons tout d'abord une méthode de modélisation qui permet de combiner les hypothèses d'opinions et thématiques. Les expérimentations que nous avons menées montrent la pertinence du mécanisme que nous proposons. Notons tout de même qu'il n'existe à ce jour aucun protocole permettant d'évaluer réellement la pertinence d'une modélisation ou encore de comparer les différentes représentations de l'image de marque que l'on peut générer. Nous montrons toutefois que la prise en compte de ces informations permet une analyse automatique plus précise des contenus, avec une robustesse accrue en présence de contenus bruités. Toutefois, l'intérêt principal de la modélisation que nous proposons est qu'elle permet de proposer pour un non initié des éléments plus intelligibles que les obscures mesures de performances, de systèmes automatiques, prisées par les chercheurs notamment dans le cadre des campagnes

7. Contenus représentant un événement, une thématique ou un utilisateur.

d'évaluation. Avec son mécanisme itératif automatique, la méthode permet également de réduire les temps de traitement (il n'est plus nécessaire de définir manuellement tous les modèles à tester, l'algorithme permet de faire émerger le modèle le plus représentatif des données) et facilite l'interprétation des hypothèses des systèmes automatiques en fournissant des descriptions claires et précises, tant sur le plan sémantique que hiérarchique. Cette méthode permet également de hiérarchiser les contenus textuels en fonction d'un critère d'intérêt défini par l'analyste ou déduit du modèle.

Nous proposons ensuite une seconde méthode de hiérarchisation des contenus textuels cette fois-ci par rapport à leur centralité au regard d'un corpus de documents. Ce corpus peut être l'ensemble des documents rattachés à une thématique ou l'ensemble des contributions d'un utilisateur. Nous montrons durant nos évaluations que la méthode d'extractions de contenus que nous proposons permet à partir de mesures statistiques classiques de hiérarchiser ces derniers afin de n'en sélectionner que la frange la plus représentative ou la plus pertinente tout en maintenant un niveau de performance similaire dans le cadre d'expériences menées à partir de collections de données réduites via notre méthode. Cette observation est d'autant plus intéressante que les volumes toujours croissants de données, générées par les utilisateurs, imposent des temps de traitement toujours plus long.

1.2 Organisation de la thèse

Maintenant que nous avons délimité les questions qui sous-tendent cette thèse, nous allons les explorer chacune à la suite des autres. Nous présentons notre travail de la manière suivante : nous consacrons un chapitre à chacune des problématiques évoquées précédemment. Nous commençons par développer un état de l'art des travaux visant à traiter de manière automatique les contenus publiés sur les différents médias sociaux dans le chapitre 2. Puis, nous présentons dans le chapitre 3 les jeux de données que nous avons utilisés et les protocoles d'évaluation que nous avons suivis tout au long de cette thèse. Le chapitre 4 présente une première contribution, qui est un protocole d'analyse des contenus textuels traitant de l'image de marque d'entités. Ce chapitre reprend les recherches que nous avons publiées dans (Cossu et al., 2014, 2015a). Dans le chapitre 5, nous proposons une méthode qui permet notamment d'identifier et caractériser les membres jouant des rôles influents dans la propagation des opinions et de l'image de marque dans les médias sociaux, le tout à partir de leurs écrits. Ces travaux sont issus d'une version revue et enrichie des travaux publiés dans (Cossu et al., 2015b). Enfin, le chapitre 6 est consacré à la modélisation de l'image de marque et la présentation de l'information dans les

médias sociaux. Nous y présentons notre dernière contribution, cette dernière vise à évaluer l'apport des concepts modélisés par la méthode précédente pour la compréhension de l'image de marque. Nous introduisons un modèle simple et nous explorons sa pertinence. Nous nous intéressons également à des méthodes de sélection de messages afin de réduire la complexité des calculs étant donné les masses de données que nous sommes amenés à manipuler. Il s'agit d'une version étendue des travaux présentés dans (Cossu et al., 2015b,?). Pour finir, le chapitre 7 clôt cette thèse en récapitulant nos principales observations, tout en proposant plusieurs pistes de poursuites des travaux de recherche.

Chapitre 2

Messages postés sur les médias sociaux, une nouvelle mine d'informations

Sommaire

2.1 Quels usages ?

Les médias sociaux permettent aux utilisateurs de partager, quelque soit la plate-forme, divers types de contenus notamment ceux inclus dans les messages textuels. Les aspects les plus étudiés sont ceux qui concernent les aspects sociaux entre les individus et tout ce qui est rattaché à ces contenus publiés. Dans le premier cas, c'est principalement sur l'utilisateur, son profil et ses relations, que l'on se concentre. Dans le deuxième cas, l'objectif est de chercher à comprendre le sens de ces contenus et trouver des applications auxquelles ils seraient utiles. C'est ce second cas qui nous intéresse ici. Notre cadre applicatif, l'analyse des images de marque sur le Web 2.0 nécessite d'être en mesure de prendre en compte le contenu textuel des messages, avis, et autres commentaires déposés par les entités étudiées ou par les internautes. On trouve dans la littérature beaucoup de publications centrées sur l'analyse des contenus de messages issus de réseaux sociaux et particulièrement Twitter.

Les messages publiés sur les médias sociaux étant régis par leurs propres règles et styles, il a d'abord fallu de nombreuses études pour proposer des systèmes de normalisation et correction de texte. Citons, par exemple, l'Atelier d'Analyses des Contenus du Web 2.0 (CAW 2.0) [1] dont l'objectif est entre autres de produire des versions corrigées des messages. Par la suite, beaucoup d'autres ateliers du même genre ont émergé et abouti à la création de nombreuses collections de données permettant à tout un chacun de construire et évaluer son propre système sur de nombreuses tâches. Cependant, il n'existe toujours pas de ressources permettant d'évaluer des problèmes plus spécifiques posés par ces messages comme l'analyse d'image de marque. Là, un travail plus conséquent d'annotation est nécessaire pour réellement comprendre les intentions des auteurs de ces messages.

Les travaux de (Jansen et al., 2009) marquent peut-être un tournant lorsque ces derniers affirment que l'on peut considérer le média (ou réseau) social Twitter comme l'équivalent d'un bouche-à-oreille numérique où 19% des messages seraient directement adressés à des services clients avec une forte tendance positive (50% contre 33% de messages négatifs). Jansen *et al.* ont également proposé un système automatique qui permet de suivre les évolutions dans le temps d'opinions au sujet d'une marque sans toutefois en proposer d'analyse fine. Ces travaux font alors de Twitter, mais aussi des autres médias sociaux, une nouvelle mine d'information qui attire particulièrement les chargés de *veille relationnelle* de grands groupes commerciaux.

1. http://caw2.barcelonamedia.org/

Dans le même temps, des études comme celles de (Mascaro et al., 2012) ou encore (Park et al., 2011) proposent de suivre l'expérience du discours politique en ligne en analysant les réactions d'utilisateurs, l'orientation des nouvelles et réactions des utilisateurs à ces dernières comme le proposent également (Sobkowicz et Sobkowicz, 2012). Dans une période électorale, c'est plus spécifiquement à la majorité silencieuse que l'on s'intéresse au travers des éléments qui sont absents de ces discours numériques. D'ailleurs, les élections allemandes de 2009 (Tumasjan et al., 2010) et américaine de 2010 (Livne et al., 2011) ont été marquées par de forts progrès des méthodes automatiques, ces résultats étant remis en question plus tard par (Jungherr et al., 2012) et (Metaxas et al., 2011). Le « *blogging* » est depuis devenu un outil fiable pour prédire les résultats du box office (Sadikov et al., 2009) ou les tendances de la bourse (Bollen et al., 2011). Dans un tout autre registre, (Sadilek et al., 2012) proposent de modéliser la propagation d'une épidémie, (Sakaki et al., 2010) souhaitent eux affiner la détection de phénomènes naturels comme les tremblements de terre, le tout, à partir des réactions observées à ce sujet sur Twitter. Les débouchés de ce genre d'études permettent d'aider les services de secours lors de catastrophes naturelles, chacun étant en mesure de communiquer publiquement la situation dans sa ville, son quartier. Cependant, certaines personnes mal-intentionnées profitent de cette situation en usurpant des comptes officiels dans le but d'annoncer de fausses nouvelles afin de créer des réactions en chaîne [2].

Plus en phase avec nos objectifs, de nombreux travaux de recherche, à la croisée du Traitement Automatique des Langues, des statistiques textuelles et de la fouille de données se penchent sur le problème de la détection d'opinions et cherchent à automatiser au maximum la détection et l'analyse des sujets (ou thématiques) traités dans des corpus de données. Ces méthodes s'appuient principalement sur les informations fournies par le document en se basant sur la présence et les occurrences de certains mots voire d'expressions (suites de mots) ou plus spécifiquement d'entités nommées. Cela permet sans se restreindre à un domaine particulier d'effectuer une analyse mais cette dernière est par contre dépendante des formes (vocabulaire et style) présentes dans le corpus traité. Quelques travaux exploitent la cohésion lexicale ; il s'agit alors de s'appuyer sur un réseau lexical regroupant les mots proches. Ce réseau nécessite l'ajout de connaissances extérieures spécifiques à une langue, en utilisant par exemple des dictionnaires, des thésaurus, ou encore des bases de données lexicales comme WordNet (Miller, 1995). Enfin, il est également possible de détecter des documents parlant de sujets similaires sur le fond, mais empruntant une forme différente, en utilisant la structure des documents (comme pour les travaux de détection de citations). Ces approches sont par contre sensibles à la qualité de langage et ne permettent pas de prendre en compte l'apparition et l'utilisation

2. L'annonce par un grand journal de la mort du président américain Obama dans un attentat à la maison blanche a été suivie de fortes baisses des bourses internationales.

d'expressions spécifiques ainsi que leurs évolutions rapides dans les médias sociaux.

Jusqu'à il y a peu encore, la plupart de ces travaux ne se concentraient que sur les analyses d'opinions associées à des produits. Ces analyses concernaient surtout les sites de commerce électronique (hôtels, restaurants, produits électroniques ou biens culturels). D'une part, ceci était dû au fait que ces données étaient facilement exploitables[3]. D'autre part, ces sites avaient besoin de ces analyses afin de nourrir leur système de recommandation pour suggérer des produits à d'autres utilisateurs et ainsi augmenter leurs ventes (Gaillard, 2014).

Toutefois, sur la plupart de ces sites, s'il n'y a pas d'ambiguïté, un commentaire concernant un bien matériel tel qu'un téléphone ne concerne que le produit en question. L'avis de l'utilisateur est connu vu que ce dernier a noté le produit sur une échelle de 1 à 5 étoiles (échelle commune à la plupart des plate-formes de commerces). De plus, les concepts généralement associés à un bien matériel ne sont pas les mêmes que ceux que l'on associe à une oeuvre telle qu'un film. Enfin, concernant justement les biens culturels, les avis rédigés sur les plate-formes de commerce, et c'est plus généralement vrai avec les médias sociaux, ne prennent pas du tout la même forme et n'utilisent pas le même vocabulaire que les avis présents sur les sites et blogs plus spécialisés ce qui limite le portage des méthodes entre les différents types de médias (Mostafa, 2013).

Dans le cadre d'une analyse d'image de marque, être à même de pouvoir identifier les différents aspects sur lesquels sont exprimées les opinions est reconnu comme étant un problème difficile. Et ceci d'autant plus quand cette analyse est effectuée à partir de messages bruités et de faible longueur comme les tweets (Amigó et al., 2013a).

2.2 Quelles méthodes pour analyser les contenus ?

2.2.1 Représentation des contenus

La plupart des méthodes d'analyse de contenus textuels se basent sur une notion de similarité. Il n'est pas possible à partir des textes dans leur forme naturelle de déterminer la valeur de ces similarités. Afin de permettre le traitement automatique des contenus par des systèmes, nous avons besoin de traduire ces documents dans une représentation intermédiaire que la machine est alors capable de manipuler. Les travaux de (Salton et al., 1975) posent les bases

3. La définition des caractéristiques définissant un téléphone mobile pouvant être généralisée à l'ensemble des modèles de téléphones mobiles.

de cette notion au travers de la représentation vectorielle des textes. Dans cette représentation, chaque document est un sac de mots (unigrammes) auxquels une pondération a été associée. Cette représentation suppose l'indépendance de l'apparition des termes les uns par rapport aux autres. Cette règle d'indépendance s'est relâchée avec l'utilisation de n-grams (suite d'unigrammes ordonnés) qui a permis d'améliorer l'efficacité des systèmes (Salton et al., 1974). (Metzler et Croft, 2005) vont par la suite aller plus loin en intégrant justement les dépendances entre les apparitions de certains mots. Les chercheurs en Traitement Automatique des Langues ont continué à développer des méthodes plus sophistiquées tenant compte des spécificités de chaque langue (Smeaton, 1999).

2.2.2 Méthodes d'analyses

Nous continuons cette section en rappelant brièvement en quoi consiste la tâche de catégorisation de textes (également appelée, classification de documents) du point de vue des domaines de l'apprentissage automatique (Mitchell, 1999; Theodoridis et Koutroumbas, 2009) et de sa variante par transfert [4] (Pan et Yang, 2010).

Beaucoup de travaux de catégorisation de textes reposent sur l'identification d'une catégorie à laquelle un document est rattaché. Ces catégories sont généralement prédéfinies clairement et possèdent une sémantique distincte. La littérature du domaine de la catégorisation de textes s'intéresse aux notions de zones de documents, ces zones pouvant être des paragraphes ou parfois des phrases. Toutefois, travailler directement sur des documents de type « *Micro-Blogs* », c'est-à-dire des documents dont la taille est généralement inférieure à celle d'une phrase, réduit la granularité de notre étude. Nous limiterons donc cet état de l'art aux travaux de portée similaire à nos contributions soit : des travaux de catégorisation de mono-étiquette de « *Micro-Blogs* », sans structure complexe entre les catégories. Nous pouvons encore distinguer plusieurs groupes de travaux autour de cette problématique de catégorisation :

— ceux généralement basés sur des approches supervisées visant à catégoriser des textes selon une catégorie d'opinion globale (positive, négative et encore trop rarement neutre) ou catégorie d'aspects (il est possible de recourir à des bases de connaissances d'aspects pré-établies) ;
— ceux qui à l'inverse utilisant des versions améliorées des mêmes approches visent à catégoriser seulement une partie d'un texte parfois même un aspect précis. Ces travaux imposent que l'entité qui fait l'objet de l'analyse (par exemple, un produit, une personnalité, une entreprise...

4. L'apprentissage par transfert à pour objectif de reconnaître et appliquer des connaissances apprises à partir de tâches antérieures similaires à celle que l'on souhaite traiter et pour laquelle nous ne disposons pas encore de données.

) ait préalablement été décomposée en plusieurs concepts que l'on souhaite pouvoir analyser séparément. Pour rester pertinente, cette décomposition doit se faire en collaboration avec les chargés de communications de l'entité. Ces travaux visent donc à une analyse fine des opinions exprimées en langue naturelle ;
— ceux visant à récupérer un élément précis d'information visant à trouver le marqueur particulier d'une opinion donnée quel que soit l'aspect ;
— ceux, utiles aux précédents, dont l'objectif est la désambiguïsation.

2.2.3 Méthodes d'apprentissage automatique

Il existe selon la littérature plusieurs techniques d'apprentissage automatique. Nous pouvons distinguer deux branches : l'apprentissage appelé « *dans le domaine* » et celui dit « *hors domaine* » qui fait référence aux méthodes de transfert de connaissances. Ces méthodes comportent trois grandes familles d'approches que l'on peut considérer en fonction des données dont on dispose, celles qui ne sont pas supervisées, celles qui le sont et enfin les approches dites « *semi-supervisées* ». Ce qui caractérise les approches supervisées c'est la présence de grandes quantités de connaissances (sous la forme exemples annotés). Elles sont limitées par ces quantités et la disponibilité des données. La tâche est alors confiée aux méthodes semi-supervisées et non supervisées qui sont nourries avec des masses de données non-annotées que l'on obtient facilement. Ces dernières méthodes comportent souvent les algorithmes de « *regroupement* », dont l'objectif est de partitionner des données en différents groupes en se basant principalement sur des similarités à l'intérieur des données. Elles sont généralement appliquées lorsque l'on ne dispose de pas ou très peu de connaissances sur les données afin dégager des sous-groupes pouvant faire l'objet de traitements particuliers.

Revenons plus en détails sur les méthodes semi-supervisées aussi qualifiées de « *faiblement supervisées* » ou à « *bases d'amorces* », leurs applications se rapprochant de notre cas d'étude. Ces méthodes essayent de tirer parti d'un ensemble limité d'exemples annotés pour propager ces annotations sur un ensemble important de documents non annotés. La littérature distingue plusieurs variantes dont nous citons les trois plus utilisées :
— l'auto apprentissage (Zhu, 2005) l'objectif est d'entraîner un système de catégorisation à reproduire au mieux les annotations d'un petit ensemble d'évaluation ;
— le co apprentissage (Blum et Mitchell, 1998), vise à entraîner conjointement plusieurs systèmes indépendants et itérer jusqu'à ce qu'ils arrivent

à obtenir des performances complémentaires [5] ;
— l'apprentissage actif (Settles, 2012). contrairement aux deux précédents, ne requiert que peu d'annotations pour l'entraînement, l'objectif étant de sélectionner les meilleurs exemples d'entraînements permettant de retrouver l'annotation de l'ensemble des documents non retenus.

Nous avons également proposer l'étude d'un modèle que nous avons appelé « *flip-flop* » (Cossu et al., 2014). Ce dernier consiste à croiser ces variantes en combinant plusieurs systèmes automatiques, entraînés sur un ensemble restreint de donnés, dans le but d'annoter de façon fiable un ensemble non-annoté. Les « *meilleures* »annotations automatiques sont alors sélectionnées pour être introduite dans l'ensemble d'entraînement ou pour être valider par un expert. Ce processus est répété itérativement de manière à réduire la taille de l'ensemble non-annoté.

2.2.4 Enrichir les tweets, une solution ?

Nous avons déjà évoqué précédemment la possibilité d'utiliser des systèmes de reconnaissance d'Entité Nommées pour approfondir le traitement des contenus (Derczynski et al., 2015; Khalid et al., 2008; Laniado et Mika, 2010). Il est d'ailleurs fréquent d'utiliser de tels systèmes dans des tâches de désambiguïsation ou l'extraction de données biographiques (nom, prénom, courriel, adresse, etc.).

D'autres travaux comme ceux de (McDonald et al., 2015; Qureshi et al., 2014) se réfèrent à d'autres types de ressources additionnelles pour enrichir le contenu des tweets afin d'améliorer les performances de leurs systèmes. Ces approches sont toutefois soumises à évolutivité et la disponibilité des ressources qu'elles utilisent (Wikipédia notamment) (Meij et al., 2012). De plus, en raison de leur dépendance vis-à-vis du langage et du domaine, elles pèchent par leur manque de portabilité. La plupart d'entre elles sont bâties à partir d'une connaissance introduite manuellement et dépendent donc du domaine d'application et des données auxquelles elles sont confrontées (Weerkamp et al., 2009).

Les approches de (Koolen et Kamps, 2010) et (Metzler et al., 2009) proposent d'utiliser les liens hyper-textes présents dans les documents et d'extraire du contenu additionnel à partir des pages liées aux documents. Une autre approche plus ancienne consiste à enrichir les documents à partir de la collection elle-même (Croft et Harper, 1979; Ruthven et Lalmas, 2003; Salton, 1971; Salton et Buckley, 1997). Dans les deux cas, une sélection des N meilleurs termes les plus importants est utilisée pour enrichir les documents.

5. Certains algorithmes peuvent être plus adaptés à une catégorie ou un type de documents.

2.2.5 Caractéristiques des documents et marqueurs intrinsèques

Il est admis par une partie de la communauté que l'on peut distinguer des documents de type Micro-Blogs portant une opinion positive ou négative en fonction de leur longueur, du nombre de noms ou verbes voire de la présence d'un marqueur de négation. Ces hypothèses sont d'autant plus vraies lorsque l'on doit faire face à des opinions très marquées exprimées brièvement(Cossu et al., 2013). Citons à ce titre les récents travaux de (Damak et al., 2013) basés sur la recherche et la sélection de caractéristiques d'un document dans le but de trouver des marqueurs spécifiques à chaque catégorie. Ce genre d'approches discrédite alors totalement le contenu textuel pour modéliser les messages. Comme nous souhaitons justifier notre analyse auprès de l'expert, ces méthodes n'emportent pas nos suffrages. De plus il n'est pas envisageable d'imaginer l'utilisation des mêmes méthodes lorsque l'on s'intéresse à la catégorisation de documents selon différentes thématiques. En effet, in message portant sur l'éthique peut *a priori* avoir la même taille ou les mêmes caractéristiques qu'un autre message qui porte sur l'innovation ou les performances. Néanmoins, sur des documents beaucoup plus longs, comme de longs billets de blogs, ces méthodes ne présentent pas ou peu d'intérêt.

Il existe dans la littérature d'autres méthodes combinant différents descripteurs. Citons par exemple la méthode développée par (Bonnefoy et al., 2013) dans le cadre du challenge «*Knowledge Base Acceleration*» (KBA) dans la conférence TREC 2012 (Frank et al., 2012). Cette campagne d'évaluation ressemble fortement à la question de la détection de niveau de priorité de RepLab 2013. Le challenge KBA se rapporte au filtrage temporel de collection de documents plus ou moins pertinents par rapport à une liste de 29 entités choisies sur Wikpiédia. Ce filtrage se matérialise par l'attribution d'un niveau de priorité (ou centralité) par rapport à l'entité, alerte, important, sans importance, sans rapport avec l'entité. Si les définitions des problèmes sont similaires, la différence principale réside dans le type de documents traités, de longues pages web ou des billets de blogs et des Micro-Blogs très brefs.

La méthode de (Bonnefoy et al., 2013) repose sur l'identification des caractéristiques intrinsèques des documents pertinents au travers de trois types de caractéristiques : (i) les caractéristiques intrinsèques des documents, (ii) le profil de l'entité et (iii) la temporalité (Bonnefoy et al., 2013). Chaque document (message) est donc représenté par cet ensemble de caractéristiques qui est combiné par un système de classification basé sur des forêts aléatoires. Le système doit, au regard des caractéristiques, déterminer à quel point le message est pertinent pour l'entité. Le tableau 2.1 donne un exemple de l'ensemble des caractéristiques retenues par cette approche.

$TF(e,d)$	Fréquence du terme e dans le document d
$TF_{10\%}(e,d)$	Fréquence du terme e dans chaque dixième de d
$TF_{20\%}(e,d)$	Fréquence du terme e dans chaque cinquième de d
$C(sent,e,d)$	Nombre de phrase contenant le terme e
$entropy(d)$	Entropie du document d
$length(d)$	Nombre de termes dans le document d
$SIM_{1g}(d,sd)$	Similarité cosinus entre d et l'article Wikipédia de l'entité, (sur la base des uni-grams)
$SIM_{2g}(d,sd)$	Idem sur la base des bi-grams
$TF(re,d)$	Fréquence du nom canonique de l'entité dans le document d
$TF(reL,d)$	Mentions d'autres entités (y compris encapsulés dans les liens) dans le document d
$TF(e,d).IDF(e,1h)$	Fréquence (et fréquence inverse) du terme e dans le document d en une heure dans la collection
$DF(e,1day)$	Nombre de documents de la collection mentionnant l'entité e ce jour
$DF(e,7d)$	Nombre de documents de la collection mentionnant l'entité e cette semaine
$Var(DF(e,7d))$	Variance de ce nombre de document sur 7 jours
$TF(e,7d)$	Fréquence du nom canonique de l'entité sur 7 jours
$TF(e,title,7d)$	Fréquence du nom canonique de l'entité dans les titres sur 7 jours

TABLE 2.1 – *Caractéristiques intrinsèques des documents, profil de l'entité et caractéristiques temporelles. A noter, quand cela est possible, les fréquences sont normalisées avec la taille des documents (Bonnefoy et al., 2013).*

2.3 Modélisation et prédictions automatiques

Notre travail nécessite l'extraction précise des opinions exprimées sur les différentes facettes d'un thème donné afin de relier ces opinions aux facettes sur lesquelles elles sont exprimées. Il nous était difficile d'envisager d'utiliser des dépendances syntaxiques ou des ressources spécifiques pour notre application (celles-ci existant trop peu en français). L'objectif nous impose de présenter à l'analyste un argumentaire humainement intelligible associé à la décision que le système a prise pour un document donné. Cela nous a amené à écarter toutes les méthodes basées sur l'extraction d'une caractéristique particulière telle que la longueur du message pour justifier la décision de notre système automatique.

Nous avons appliqué des méthodes à la croisée des chemins entre la Recherche d'Information et Traitement Automatique des Langues. Ces méthodes consistent à utiliser la distribution statistique des mots et expressions pour déterminer leur rattachement à une catégorie, qui peut être une opinion ou une thématique. Cela peut être considéré comme une extension des méthodes basées sur la présence de ces marqueurs comme dans les méthodes historiques de recherche d'information. Avec ces dernières, les termes sont représentés par un booléen, soit ils sont présents dans le document soit en sont absents, peu importe leur fréquence ou leur importance. L'arrivée de la pondération statistique (Salton et al., 1975) a permis d'affiner le pouvoir discriminant de ces marqueurs. Plus que la présence ou non de termes connus, nous misons sur l'importance de ces derniers en utilisant différentes pondérations statistiques.

Pour atteindre les objectifs que nous nous sommes fixés, il nous a fallu faire quelques hypothèses, choisir une palette de méthodes appropriées, élaborer une chaîne de traitement que nous allons décrire, développer des programmes qui mettent en oeuvre les méthodes retenues et enfin mener un grand nombre d'expériences qui nous ont permis de valider ou d'invalider certaines de nos hypothèses. Nous abordons dans cette partie les différentes approches que nous avons employées conjointement pour ajuster les paramètres et optimiser les performances de nos systèmes.

2.3.1 Représentation des documents

L'indice le plus souvent employé en recherche d'information pour quantifier la similarité entre 2 documents, ou entre un document et une requête est l'indice cosinus dont la formule est donnée pour rappel (2.1). Cet indice est aussi utilisé en classification textuelle pour estimer la similarité entre un document d et une classe c. Les documents de type Micro-Blogs ayant la taille d'une phrase moyenne, nous ne nous attarderons pas ici, sur les problèmes liés à la dispro-

portion entre les différentes tailles de documents. Les cas d'études étant au plus près d'un scénario d'utilisation de production nous subissons les disproportions de taille entre l'ensemble de classes sans chercher à artificiellement les réduire pour faciliter les processus d'entraînement. Les paragraphes de billets de blogs seront analysés à cette échelle de la phrase. Forcer l'équilibre entre la taille des différentes classes établirait un biais pour l'analyse de l'image de marque.

$$cos(d,c) = \frac{\sum\limits_{i \in d \cap c} \omega_{i,d} \times \omega_{i,c}}{\sqrt{\sum\limits_{i \in d} \omega_{i,d}^2 \times \sum\limits_{i \in c} \omega_{i,c}^2}} \tag{2.1}$$

Nous le comparons à l'indice de jaccard non binaire dont la formule est donnée pour rappel (2.2).

$$jac(d,c) = \frac{\sum\limits_{i \in d \cap c} \omega_{i,c}}{\sum\limits_{i \in c} \omega_{i,c}} \tag{2.2}$$

En complément de ces mesures de similarité, nous utilisons une méthode qui repose sur une autre exploitation de la base des connaissances annotées : la méthode des K-plus proches voisins (notée KPPV et KNN par la suite). A partir des similarités calculées entre l'ensemble des documents, cette méthode permet de définir des fonctions ayant pour objectif de retrouver (et/ou regrouper) les documents les plus similaires. Cette technique est très régulièrement appliquée dans les problèmes de recommandation pour comparer différents objets et utilisateurs entre eux. Le fonctionnement de cette méthode est on ne peut plus simple. Tous les documents à analyser sont comparés à ceux qui font partie de l'ensemble d'apprentissage. Pour prendre une décision sur un nouveau document, ne seront retenus que les K documents annotés plus proches (k plus proches voisins) selon les indices de similarité (dans notre cas : jaccard et cosinus). Ces voisins votent proportionnellement à leur similarité pour la catégorie à laquelle ils appartiennent. De nombreuses fonctions heuristiques sont également considérées dans la littérature pour affiner le fonctionnement de cette méthode. Dans le cas d'une approche itérative, ces K plus proches peuvent être mobiles. L'avantage et à la fois l'inconvénient de cette méthode repose sur sa complexité. En effet, durant la phase d'apprentissage, le temps de calcul de la similarité entre l'ensemble des documents est proportionnel au nombre de documents constituant la collection de travail.

Enfin, il serait difficile de conclure cette section sans mentionner les incontournable machines à vecteurs de support ou séparateurs à vaste marge (en anglais «*Support Vector Machine*», *SVM* (Vapnik et Vapnik, 1998)) sous l'implémentation Multi-Class [6] (Crammer et Singer, 2002). L'efficacité de cette méthode n'est plus à démontrer y compris dans les tâches difficiles de TAL (Joachims, 1998; Manning et Schütze, 1999). Nous nous sommes restreint à entraîner un classifieur linéaire en laissant les paramètres par défaut, une configuration également utilisée dans la littérature (McDonald et al., 2015).

2.3.2 Représentation des termes

Comme le montrent les formules (2.5) et (2.4), les poids $\omega_{i,x}$ sont généralement obtenus en faisant le produit de 2 facteurs : le nombre de fois (TF noté ici TF_i, x) où le terme i apparaît dans le segment x ($x = d$ ou $x = c$), et une fonction inverse (IDF) du nombre de segments ($DF_{\mathbb{C}}(i)$) parmi l'ensemble des N segments, où le terme i apparaît au moins une fois. TFi, x est parfois considérée dans sa valeur absolue ou plus généralement sous une forme normalisée (avec une échelle logarithmique par exemple). Cette représentation statistique est issue des travaux en recherche d'information de (Sparck Jones, 1972) et (Salton et al., 1975). Cela permet d'associer à chaque terme un vecteur de poids pour chaque catégorie. Nous proposons d'enrichir la formule de calcul des poids en y ajoutant un facteur discriminant Gi et en soumettant à une élasticité plus ou moins grande chacun des trois facteurs via une élévation à une puissance variable.

Le facteur discriminant G du mot i est défini comme suit (2.3) :

$$G_i = \sum_{c \in \mathbb{C}} \mathbb{P}^2(i|c) = \sum_{c \in \mathbb{C}} \left(\frac{DF_i(c)}{DF_{\mathbb{T}}(i)} \right)^2 \tag{2.3}$$

Où C est l'ensemble des catégories, $DF_{\mathbb{T}}(i)$ le nombre de documents de l'ensemble d'apprentissage $\mathbb{T}$ contenant le mot i et $DF_i(c)$ le nombre de documents dans l'ensemble d'apprentissage appartenant à la catégorie c et contenant le mot i. Ce facteur discriminant est utilisé pour pondérer la contribution $\omega_{i,d}$ du mot i dans le document d tel que (2.4) :

$$\omega_{i,d} = TF_{i,d}^{\alpha} \times log(\frac{N}{DF_{\mathbb{C}}(i)})^{\beta} \times G_i^{\gamma} \tag{2.4}$$

6. http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html

Où N est le nombre de documents dans l'ensemble d'apprentissage. Les exposants α , β et γ permettent de donner plus ou moins d'importance à chacune des métriques considérées (nous proposons et détaillons dans la section 2.3.3 une méthode permettant de fixer ces trois valeurs de manière optimale).

G est également utilisé pour pondérer la contribution $\omega_{i,c}$ de chaque mot i dans la classe c selon (2.5) :

$$\omega_{i,c} = DF_{i,c}^{\alpha} \times log(\frac{N}{DF_{\mathbb{C}}(i)})^{\beta} \times G_i^{\gamma} \tag{2.5}$$

Les SVM sont appris en utilisant la représentation vectorielle du sac-de-mots de chaque document d selon (2.6) :

$$\omega_i = DF_{\mathbb{T}}(i)^{\alpha} \times log(\frac{N}{DF_{\mathbb{T}}(i)})^{\beta} \times gini_i^{\gamma} \tag{2.6}$$

Le problème de cette représentation est que les termes qui ne sont jamais apparus dans une catégorie se voient attribuer un poids nul dans celle-ci. La mesure de similarité ignorant alors ce terme et de fait augmente la probabilité de se tromper. La littérature s'est alors intéressée sur l'attribution d'un poids même minime à ces événements non vus, cette pratique est appelée *lissage* (Chen et Goodman, 1996; Losada et Azzopardi, 2008; Zhai et Lafferty, 2001a, 2004). Les méthodes de lissage les plus populaires sont Jelinek-Mercer et Dirichlet (Chen et Goodman, 1996; Zhai et Lafferty, 2001b). Leur objectif est de re-estimer une pondération équitable pour redonner de l'importance à ces termes ignorés sans pour autant que leur influence sur la décision finale ne dépasse celle des termes bien présents dans le document.

2.3.3 Apprentissage et optimisation par tirage aléatoire

Les trois exposants qui figurent dans les formules (2.4), (2.5) et (2.6) sont autant de paramètres sur lesquels il est possible de jouer pour tirer parti d'une métrique ou réduire son impact dans le cas où l'on considère que l'une ou l'autre des mesures n'apporte pas la quantité d'information souhaitée. Il convient d'estimer leur valeur sur un corpus de développement. Cette proposition intègre donc, comme un cas particulier, le calcul des poids généralement employés en recherche d'information. Notons que l'on peut enrichir ce jeu de paramètres en différenciant un jeu dédié à d et un autre « *dédié* » à c.

L'approche classique avec des ensembles d'entraînement (également appelé ensemble d'apprentissage), de développement et d'évaluation revient à suivre le protocole suivant :

1. des exposants sont choisis aléatoirement ;
2. un modèle est appris à partir des données d'apprentissage ;
3. la pertinence du modèle est vérifiée sur l'ensemble de développement ;

Le processus étant répété un certain nombre de fois jusqu'à obtenir une pertinence maximale stable sur l'ensemble de développement. Une fois ce processus d'apprentissage terminé, on évalue ce modèle optimisé sur l'ensemble d'évaluation :

1. un modèle est appris à partir des données d'apprentissage et de développement ;
2. le tout est évalué sur l'ensemble d'évaluation ;

Toutefois, les exposants et le modèle se sont alors adaptés à l'ensemble de développement ; cela entraîne un risque de « *sur-apprentissage* ». Pour éviter ces problèmes sur des collections de données figées une pratique commune est de tirer des éléments de manière aléatoire pour constituer plusieurs lots d'évaluation (pratique appelée cross-validation) plutôt que de constituer un ensemble d'évaluation à partir de la chronologie des données. Dans le cas du suivi dynamique de l'image de marque, cette pratique n'est pas envisageable. D'une part car nous ne sommes pas censés avoir à notre disposition ces données postérieures et d'autre part car cela reviendrait à prédire des éléments d'un passé plus ou moins lointain avec des éléments qui correspondraient au « *futur* ».

Afin d'assurer une certaine réactivité de nos systèmes ainsi que de leur capacité et efficacité dans une application en temps-réel, au lieu de recourir à cette pratique de tirage aléatoire et nous proposons de respecter la dimension chronologique des données tout en tirant parti des flux de messages qui nous parviennent et que nous devons traiter à chaque instant.

2.4 Prédictions multiples et prise de décision

Nous venons de citer dans les sections précédentes quelques unes des méthodes les plus souvent utilisées pour catégoriser automatiquement des contenus textuels. Ces N méthodes disponibles peuvent être combinées de différentes manières, en passant outre les questions de votes majoritaires (sans tenir compte de la confiance estimée par le système) la méthode de combinaison la plus simple est la combinaison linéaire. Cette dernière peut-être pondérée ou non[7].

7. Les pondérations pouvant être déterminées manuellement en fonction des résultats sur un ensemble de développement ou apprises automatiquement en utilisant le processus vu dans

2.4.1 Combinaison linéaire

Considérons un système de catégorisation automatique j qui associe chaque document d de l'ensemble d'évaluation à chaque catégorie C_k avec un score de confiance $s_j(T, C_k)$ $(j = 1, ..., N)$. La catégorie c retenue est celle qui respecte la règle suivante :

$$c = \arg\max_k \left(\sum_{j=1}^{N} s_j(d, C_k) \right) \tag{2.7}$$

2.4.2 Optimisation multicritères

Nous proposons également d'utiliser les méthodes d'optimisation multicritères « *ELECTRE* » et « *PROMETHEE* » décrites par (Roy, 1991). Ces méthodes ont été utilisées dans des contextes industriels (Lamontagne et Abi-Zeid, 2006; Gourion et Josselin, 2012) très différents du cas qui nous intéresse : la combinaison de classifieurs. Ces méthodes d'aide à la décision ont pour objectif de sélectionner la meilleur décision en hiérarchisant les critères et les candidats. Dans notre cas, la « *meilleure* » décision est la « *meilleure* » catégorie prédite par un système automatique (qui correspond au « *candidat* ») pour ce document, les critères étant les différents systèmes.

« *ELECTRE* » se base sur une relation de « *sur-classement* » $\mathcal{S} \subset \mathbb{C} \times \mathbb{C}$ qui est défini sur l'ensemble des catégories $\mathbb{C}$. Cette relation sous-entend qu'une catégorie c est préférable à la catégorie c' si (i) c obtient un score plus élevé pour une majorité de systèmes automatiques et (ii) c' n'obtient pas des scores significativement plus élevés que c sur l'ensemble des systèmes restants. Plus précisément, pour chaque paire de catégorie c, c' une matrice de concordance $m(c, c')$ est calculée. Elle indique la proportion de systèmes où c obtient de meilleurs scores de confiance que c'. On estime alors que c sur-classe c' si (i) $m(c, c')$ dépasse un seuil (généralement fixé à $2/3$ des systèmes) et si (ii) c n'est pas dominée par c' sur le tiers de systèmes restant [8]. Ce processus permet de remplir un ensemble $\mathcal{S}$ de catégories n'étant pas sur-classées par d'autres. On mise alors sur le fait qu'il n'y aura qu'une catégorie dans cet ensemble, il peut-être vide si aucune catégorie ne ressort ou bien contenir n catégories (si nous considérons beaucoup de systèmes).

La méthode « *PROMETHEE* » se base sur une matrice de concordance. Pour chaque paire de catégories, (c_i, c_j), une matrice de coefficients m_{ij} qui correspond à la matrice de concordance $m(c_i, c_j)$ introduite plus haut. Pour chaque

la section 2.3.3

8. Le seuil de veto est habituellement fixé tel que $v = 0.5$.

catégorie c_i, il faut calculer deux sommes : $s_c(c_i) = \sum_j m_{ij}$ and $s_m(c_i) = \sum_j m_{ji}$. $s_c(c_i)$ qui mesure la tendance de c_i à dominer les autres catégories et $s_c(c_i)$ la tendance de c_i à être dominée. La catégorie c_i retenue est celle pour laquelle la différence $s_m(c_i) - s_m(c_i)$ est maximale.

Etant donné notre cadre d'application nous savons qu'il existe un besoin de hiérarchisation des décisions (reprenons l'exemple où des conseillers en communication proposent chacun des plans d'actions différents). En travaillant individuellement au niveau de chaque document ces méthodes se présentent donc comme des moyens complémentaires aux modélisations plus globales que nous proposons dans le chapitre 2. En considérant l'exemple d'un homme politique entouré par plusieurs conseillés proposant chacun un plan d'action différent, ces méthodes permettraient de choisir à l'intérieur de plan, action par action, celle qui serait la plus pertinente. Une exploitation fine des matrices de concordance permettraient de hiérarchiser chaque conseillé.

Chapitre 3

Méthodologie expérimentale

Sommaire

3.1 Introduction

Dans le chapitre précédent, nous avons donné un aperçu des différentes questions et hypothèses qui sous-tendent cette thèse. Nous souhaitons à partir, de méthodes simples d'analyse des contenus textuels, apprendre les compétences d'un analyste de réputation à partir des données et connaissances qu'il a amassées jusqu'ici pour ensuite les reproduire et les propager à de nouvelles données. Dans cette optique, les domaines du Traitement Automatique des Langues (TAL) et de la Recherche d'Information (RI) se construisent depuis maintenant plusieurs décennies sur une culture de la validation d'hypothèses par l'expérimentation et notamment par le biais de campagnes d'évaluation. Outre les données, au centre de ces campagnes et de leurs évaluations se trouvent les notions de mesures d'évaluation et de leur pertinence, ces dernières, indispensables à la compréhension du comportement d'un système de RI ou de TAL. Dans ce chapitre, nous détaillons ces notions d'évaluations et dressons un portrait des différentes campagnes d'évaluation en activité. Nous décrivons également les collections de test et les sources d'information que nous avons utilisées tout au long de cette thèse.

3.2 Evaluation

L'évaluation des performances des systèmes de catégorisation automatique de documents est un problème majeur toujours d'actualité. Cela a pour conséquence l'apparition continue de nouvelles métriques pour essayer de définir un protocole d'évaluation répondant à l'ensemble des besoins. Comme il n'est pas envisageable de pouvoir évaluer la qualité d'un système en temps réel à partir de la satisfaction de l'analyste auquel on présente les prédictions du système, nous utilisons le protocole d'évaluation classique du monde l'apprentissage automatique. Cela consiste à masquer la connaissance de la catégorie pour une partie des documents de notre collection et à comparer les prédictions du système à ces annotations. Afin d'évaluer de manière plus réaliste la qualité de ces annotations automatiques, les communautés de l'apprentissage automatique, fouille de textes et recherche d'informations ont développé une pléthore de métriques qui se répartissent en deux familles, celles qui se basent sur les relations entre les documents retournés et celles qui se basent sur la décision prise pour chaque document (la catégorie associée).

Dans le cadre du projet Imagiweb. Nous avons choisi d'aller plus loin que le classique protocole de tests unitaires par validation croisée de la qualité des annotations fournies par nos systèmes. Nous avons souhaité faire directement évaluer (par validation manuelle) sur de nouvelles données les résultats des al-

gorithmes directement par les utilisateurs finaux c'est à dire par des experts du domaine. Il s'agit du CEPEL pour les tweets politiques, d'EDF pour les blogs. De plus, nous avons décidé pour la suite, quand cela est possible de considérer des métriques provenant de ces deux familles évoquées ci-dessus pour évaluer nos propositions. Nous avons dans ce cas considérer deux scénarios de validation :

— « *Validation à postériori* » : le résultat des algorithmes est présenté à l'expert qui valide ou non l'annotation automatique ;
— « *Etiquetage à l'aveugle* » (ou masqué) : l'expert n'est pas informé du résultat des algorithmes. La comparaison est réalisée de manière indépendante afin de réduire le biais induit par la stratégie précédente.

Dans les deux cas, une annotation automatique non satisfaisante est, de fait, corrigée par l'expert. Cela fournit des informations sur l'échec des algorithmes mais aussi des données étiquetées supplémentaires pour améliorer les résultats.

Evaluation de catégorisation

Nous utilisons les métriques facilement interprétables communes aux tâches de classification et de fouilles de textes : précision, rappel ainsi qu'un F-Score (combinant les deux). Le rappel mesure la proportion de documents trouvés parmi tous les documents à retrouver alors que la précision mesure la proportion de documents correctement attribués parmi ceux qui ont été retournés [1]. La définition de ces deux mesures est formulée ci-dessous. La précision est définie par :

$$\text{Précision} = \frac{\text{Nombre de documents correctement attribués}}{\text{Nombre de documents attribués}} \tag{3.1}$$

Le Rappel par :

$$\text{Rappel} = \frac{\text{Nombre de documents correctement attribués}}{\text{Nombre de documents à attribuer}} \tag{3.2}$$

Dans le cas où un système ne traite qu'une partie de l'ensemble des documents, ces deux métriques varient de manière indépendante, généralement quand l'une augmente l'autre baisse. Enfin, leur principal inconvénient est qu'elles ne prennent pas en compte l'écart à la décision, c'est à dire à quel point le système est-il loin de la référence. De plus, dans le cas où la référence est floue comme pour notre application relative à la réputation, si la décision du

1. Ces deux métriques peuvent également se calculer sur un nombre restreint de documents on parle alors par exemple de la *Précision* à N documents.

système est fausse il serait intéressant de savoir à quel rang le système a placé la référence.

Ces métriques sont combinées dans un *F-Score* dont nous considérons les deux variantes suivantes. Le Micro F_Score (calculé comme indiqué en (3.3)) représente l'efficacité globale d'un système [2].

$$\text{micro F\_Score} = \frac{2 \times (\text{Précision} \times \text{Rappel})}{\text{Précision} + \text{Rappel}} \tag{3.3}$$

Et le Macro F_Score (calculé comme indiqué en (3.4)) donne autant d'importance à chacune des catégories c. Il indique la capacité du système à retrouver l'information sur l'ensemble des catégories quelles soient plus ou moins peuplées.

$$\text{macro F\_Score} = \frac{\dfrac{2 \times (\text{Précision}_c \times \text{Rappel}_c)}{\text{Précision}_c + \text{Rappel}_c}}{C} \tag{3.4}$$

Où C est le nombre de catégories.

Évaluation relationnelle

Du fait de l'interprétabilité des données et des notions de priorité manipulées comme l'alerte, (Amigó et al., 2013a) (en tant qu'organisateurs de RepLab) ont également proposé d'évaluer les systèmes d'analyse de réputation (et donc les participants aux tâches RepLab) en utilisant les métriques Reliability (R), Sensitivity (S) (combinées dans une F-Mesure F(R&S)) (Amigó et al., 2013b). On peut considérer ces deux métriques Reliability et Sensitivity comme étant équivalentes à Précision et Rappel des relations (Amigó et al., 2013b) avec l'hypothèse que la collection de données est régie par un ensemble de relations hiérarchiques entre documents ($<$, $>$). Les catégories existantes sont organisées comme suit : **Alerte élevée (ou positive)** $>$ **Alerte faible (ou neutre)** $>$ **Non alerte (ou négative)**. Il existe une hiérarchie plus simple pour la tâche de filtrage : **Pertinent** $>$ **Non-Pertinent**. Le problème de regroupement de message (dit de « *clustering* ») est régi par la relation suivante (sachant qu'un message

2. Cette mesure n'est toutefois pas toujours représentative de l'efficacité d'un système. Dans le cas d'une collection présentant un biais sur une ou plusieurs classes. Mettre tous les tweets dans la classe majoritaire permet d'obtenir un score élevé sans que cela n'indique une réelle capacité à résoudre le problème de catégorisation.

n'appartient qu'à un et un seul groupe) : **Cluster A** ! = **Cluster B**. L'évaluation revient à comparer les relations (par paires) proposées par le système aux références définies lors de l'annotation.

La Sensitivity est définie par :

$$\text{Sensitivity} = \frac{\text{Nombre de paires concordantes entre système et référence}}{\text{Nombre de paires existant dans le système}} \tag{3.5}$$

La Reliability est définie par :

$$\text{Reliability} = \frac{\text{Nombre de paires concordantes entre système et référence}}{\text{Nombre de paires existant dans la référence}} \tag{3.6}$$

Cette évaluation permet de s'affranchir de la prédiction d'appartenance binaire à une classe en évaluant la capacité du système à avoir détecté une relation entre deux tweets (même niveau d'alerte, ou supériorité/infériorité de l'un par rapport à l'autre). Il existe toutefois une lacune sur la précision des décisions car l'écart hiérarchique n'est pas mesuré. De ce fait, les relations suivantes sont considérées comme équivalentes :

— **Alerte élevé (ou positif)** > **Non alerte (ou négatif)** ;
— **Alerte faible (ou neutre)** > **Non alerte (ou négatif)** ;
— **Alerte élevé (ou positif)** > **Alerte faible (ou neutre)**.

Autre inconvénient, cette métrique impose d'avoir défini, au préalable, une hiérarchie entre les différentes catégories considérées. Cette notion de hiérarchie entre les catégories tomber sous le sens quand il s'agit évaluer des niveaux d'alerte ou des intensités d'opinions (polarité), elle perd toutefois son intérêt lorsque l'on s'intéresse à des thématiques où, dans ce cas, sauf à définir des relations entre les thématiques nous retombons quasiement sur une évaluation classique de catégorisation.

A noter également, dans le cadre d'une évaluation multi-entité comme c'est le cas pour RepLab et Imagiweb, ces métriques sont calculées entité par entité puis moyennées.

Évaluation de préférence

Travailler avec des données issues de réseaux sociaux implique de manipuler de grandes quantités de données, lorsque l'on s'intéresse à l'analyse de

l'image de marque on souhaite se focaliser sur certains contenus qui peuvent être considérés comme plus pertinents. Cette notion de pertinence rejoint celle considérée dans le cadre des systèmes de recherche d'informations (comme les systèmes de filtrage et recommandation). Pour évaluer leurs performances, les campagnes comme TREC (Text REtrieval Conference) ont établi une métrique standard la «*Mean Average Precision*» (MAP). Cette métrique correspond à la moyenne des précisions moyennes pour chacune des requêtes. Elle mesure les performances globales d'un système vis a vis d'une requête (une moyenne des MAP obtenues pour chaque requête est habituellement faite lorsqu'il y a plusieurs requêtes).

La MAP est définie par l'équation suivante :

$$\text{MAP} = \frac{1}{N} \sum_{i=1}^{N} \mathit{AveP(i)} \tag{3.7}$$

Avec :

$$\mathit{AveP(i)} = p(i) \times R(i) \tag{3.8}$$

où N est le nombre total document, $p(i)$ la précision au rang i (à i documents restitués) et $R(i)$ vaut 1, si le $i$ème document restitué est pertinent et 0 si le $i$ème document restitué n'est pas pertinent.

Cette métrique présente toutefois un inconvénient, les éléments qui apparaissent en tête de liste ont plus de poids que ceux qui apparaissent en queue, cela signifie qu'une erreur sur les premiers documents sera plus pénalisante qu'une erreur sur les derniers. Ce déséquilibre tombe sous le sens pour un moteur de recherche d'information où l'objectif est bien d'évaluer les qualités des premiers documents renvoyés car étant ceux que l'utilisateur sera le plus à même de consulter. Dans notre cas, il conviendrait de donner le même poids aux N documents attendus dans la classe. Ce petit inconvénient reste toutefois acceptable dans le cas où il faut ordonner des alertes ou des utilisateurs selon un niveau d'influence où la métrique est très pertinente. Cette évaluation permet toutefois de comparer le pouvoir de catégorisation des différents systèmes, catégorie par catégorie.

3.3 Données et évaluations

L'ensemble des méthodes présentées dans le chapitre 2 ont contribué à l'annotation des données relatives au projet Imagiweb. Afin de vérifier leurs per-

formances vis-a-vis de l'état de l'art, les méthodes ont également été testées sur les collections RepLab 2013 et 2014.

3.3.1 Collections Imagiweb Blogs et Twitter

Nous décrivons dans cette sous-section les collections de données Imagiweb. Ces collections sont dédiées à l'analyse de l'image d'hommes politiques et d'entreprise sur le Web 2.0 [3]. Cette collection se divise en deux parties, une première se focalisant sur les messages Twitter traitant des candidats à la dernière élection présidentielle française de mai 2012 et une deuxième partie composée de billets de blogs relatifs au leader énergétique français EDF. Les données ont été annotées dans l'objectif de répondre aux questions suivantes :

1. Est-ce que ce message a un impact sur l'image de l'entité ?
2. Existe-t-il une partie de ce message qui véhicule une opinion ?
3. Sur quel(s) aspect(s) porte l'opinion ? [4]

L'objectif est de propager la connaissance de cette amorce annotée à des quantités plus importantes de données. Une étape de sélection automatique sur ces nouveaux documents annotés permet ensuite d'établir une image de marque.

Sous-collection Twitter

Cette sous-collection comporte initialement 7,283 tweets (6,362 contenus uniques) en français récupérés à partir de mots-clés (concernant Nicolas Sarkozy et François Hollande) entre mars et décembre 2012. Ces messages ont ensuite été annotés manuellement (plusieurs fois pour certains messages, portant le nombre d'annotations à un peu plus de 12000) par un collège d'annotateurs selon les critères suivants :

— **Opinion** $\subseteq$ **tweet_id** $\times$ {(très) Positive, Neutre, (très) Négative et Ambiguë}, ces polarités ont finalement été ramenées aux trois niveaux que l'on peut habituellement trouver dans la littérature (Carrillo-de Albornoz et al., 2014; Villena Román et al., 2013) ;

3. La partie concernant les hommes politiques est disponible publiquement avec l'ensemble des annotations à l'adresse suivante : http://mediamining.univ-lyon2.fr/velcin/imagiweb/dataset.html

4. Il est rare qu'un tweet porte sur plusieurs aspects mais ce cas est relativement fréquent pour un billet.

— **Thèmatique** $\subseteq$ **tweet_id** $\times$ { Entité (pas de thème précis), Attribut (sondage ou soutien), Bilan, Compétence (expertise), Ethique, Injonction (appel au vote ou à la démission), Performance communicationnelle, Personne (caractéristique physique, vie privée), Positionnement politique, Projet}. Cette taxonomie de thématiques a été établie avec la collaboration d'experts en sciences politique. (Velcin et al., 2014) reviennent en détails sur la procédure d'annotation.

Comme le notent (Velcin et al., 2014), il existe de cas de divergence d'avis entre annotateurs. Nous développons, dans la section suivante le protocole, que nous avons retenu pour résoudre ces litiges en exploitant notamment les contenus textuels.

La collection a été découpée chronologiquement en deux parties, apprentissage (Tr) développement (les 1,100 derniers messages associés à chaque candidat), nous reviendrons plus tard sur la collection de test. Le tableau 3.1 montre la distribution des opinions et des catégories thématiques pour chaque entité dans l'ensemble de la sous-collection. Ce découpage chronologique tombe sous le sens dans notre cas d'études, il serait contraire à toute logique d'utiliser, pour entraîner nos systèmes, des données postérieures à l'évaluation.

Thématique	**Hollande**	**Sarkozy**	**Total**
Attribut	287	323	610
Bilan	135	413	548
Compétence	233	110	343
Entité	795	895	1,690
Éthique	295	458	753
Injonction	94	89	183
Performance	249	161	410
Personne	331	290	621
Positionnement	610	244	854
Projet	176	85	261
Total	3,205	3,068	6,273
Opinion	**Hollande**	**Sarkozy**	
Négative	1,907	1,626	3593
Neutre	839	839	1678
Positive	434	717	1151
Total	3,180	3,182	6,362

TABLE 3.1 – *Répartition des opinions et des catégories thématiques pour chaque entité.*

Nous utilisons pour l'évaluation une sélection de 5,216 tweets collectés tout le long de l'année 2013 (environ 430 par mois) concernant Nicolas Sarkozy et 3,705 messages récupérés aux mois de mars et avril 2013 concernant François

Hollande. Ces messages proviennent d'un ensemble de messages extraits entre janvier 2012 et décembre 2014, ensemble qui est organisé comme suit :

— 240,000 tweets (6,700 par mois) au sujet de François Hollande ;
— 81,000 messages (2,500 par mois) concernant Nicolas Sarkozy.

Les messages sélectionnés ont été annotés automatiquement avec les systèmes décrits précédemment puis vérifiés manuellement par deux experts en science politique selon le protocole d'annotation du projet Imagiweb. Notons que pour chaque entité, les messages ont été séparés en deux lots, un premier pour lequel les annotateurs devaient vérifier l'annotation automatique, un second lot pour lequel les annotateurs effectuaient une annotation à l'aveugle. Nous avons *a posteriori* comparé les annotations de l'expert aux hypothèses du système automatique. Ce second cas permet d'évaluer l'influence de la proposition du système automatique sur le choix final de l'annotateur, ce dernier pouvant par défaut se reporter au choix du système ou trouver satisfaisante l'hypothèse du système même si cette hypothèse peut-être discutable car une autre catégorie aurait pu être tout aussi juste. Enfin, il existe le cas du rejet de l'annotation automatique dans le cas où cette dernière serait fausse.

Le tableau 3.2 montre la distribution des opinions et des catégories thématiques pour chaque entité dans l'ensemble de cette collection d'évaluation.

Thématique	**Hollande**	**Sarkozy**
Attribut	551	694
Bilan	270	608
Compétence	103	37
Entité	985	1476
Ethique	523	1516
Injonction	100	79
Performance	149	91
Personne	277	411
Positionnement	569	261
Projet	78	43
Total	3,605	5,216
Opinion	**Hollande**	**Sarkozy**
Négative	2619	3123
Neutre	860	813
Positive	126	1280
Total	3,605	5,216

TABLE 3.2 – *Répartition des opinions et des catégories thématiques pour chaque entité dans l'ensemble d'évaluation de cette collection.*

Sous-collection de Blogs

Cette collection constituée de 3,300 billets de blogs francophones est centrée sur l'entreprise EDF et plus précisément sur la thématique du nucléaire [5]. C'est cette dernière qui s'est chargé de la collecte des données. Ces billets se regroupent sous trois thématiques principales, l'emploi, les tarifs et la sécurité. Les billets de blogs pouvant être très longs, le message soumis à l'analyse est un paragraphe comportant a minima 250 caractères avant et après l'occurrence de l'entité EDF. La dimension retenue (après validation par un analyste) est celle du paragraphe de 6 phrases comportant au centre la mention « *EDF* ». Pour constituer un ensemble d'entraînement pour les approches d'apprentissage automatique, une partie des blogs (600 billets uniques) a été manuellement annotée [6] par un sémiologue en collaboration avec des personnels de l'entreprise selon les critères suivants :

- **Opinion** $\subseteq$ **blog_id** $\times$ {(très) Positive, Neutre, (très) Négative et Ambigue}, comme dans le cas précédent, ces polarités ont finalement été ramenées à trois niveaux ;
- **Thèmatique** $\subseteq$ **blog_id** $\times$ {Risques, Stratégie, Tarifs} ;
- **Sous-Thèmatique** $\subseteq$ **blog_id** $\times$ {Risques :Accidents, Tarifs :Client, Tarifs :Coûts de production, Risques :Déchets, Tarifs :Démantèlement, Stratégie :Économique, Risques :Expertise, Tarifs :Loi Nôme, Stratégie :Politique, Risques :Transparence} Ces dernières étant organisées sous la forme d'une hiérarchie où la thématique conditionne la sous-thématique.

L'ensemble des billets restants (environ 2,700) a été considéré comme un ensemble d'évaluation. Cet ensemble a été annoté automatiquement puis partiellement vérifié manuellement par un expert selon le protocole décrit précédemment. Cette vérification concerne 609 billets uniques, certains d'entre eux ont reçu plusieurs annotations différentes de l'expert (soit un total de 768 annotations [7]). Le tableau 3.3 montre la distribution des opinions et des catégories thématiques pour chaque entité dans l'ensemble des documents annotés de la collection. Nous pouvons noter que la question des tarifs est une préoccupation secondaire en termes de volume (et notamment l'évolution des tarifs clients), les opinions semblent également assez tranchées avec une faible proportion de messages neutres. La question du traitement des déchets semblent étonnement peu traitée par les blogeurs.

5. Les billets ont été collectés de janvier 2011 à septembre 2012. Ils contiennent au moins une des désignations suivantes de l'entité EDF : « *EDF* », « *E.D.F* », « *Electricité de France* », « *l'électricien français* ».

6. Certains billets ont reçu plusieurs annotations, soit au total 681 annotations.

7. Les quantités totales peuvent varier étant donner que certains contenus ont pu recevoir plusieurs annotations de thématiques avec une même opinion ou vice-versa.

Thématique	**Entraînement**	**Évaluation**	
Risques	223	321	544
Stratégies	274	290	564
Tarifs	184	156	340
Total	681	767	1,448
Sous-Thématique	**Entraînement**	**Évaluation**	
Tarifs : Client	49	48	97
Tarifs : Coûts de production	76	44	120
Tarifs : Loi Nome	59	64	123
Risques : Accidents	69	126	195
Risques : Déchets	14	18	32
Risques : Démantèlement	54	49	103
Risques : Expertise	86	128	214
Stratégies : Économique	97	73	170
Stratégies : Politique	100	133	233
Stratégies : Transparence	77	82	159
Total	681	765	1,452
Opinion	**Entraînement**	**Évaluation**	
Négative	437	527	964
Neutre	86	71	157
Positive	161	170	331
Total	684	768	1,452

TABLE 3.3 – *Répartition des opinions et des catégories thématiques pour chaque entité dans les sous-collections d'entraînement et d'évaluation.*

3.3.2 Collection RepLab'2013-14 Twitter

Nous décrivons dans cette sous-section la collection de données CLEF RepLab2013-2014 [8]. Cette collection a été conçue pour une campagne d'évaluation dédiée au suivi de la réputation en ligne d'entités sur le réseau social Twitter dans le contexte des éditions 2013 et 2014 de la *Conference and Labs of the Evaluation Forum* [9] (CLEF). Le problème a été découpé par les organisateurs de RepLab (Amigó et al., 2013a, 2014) en cinq sous problématiques relationnelles :

— **Filtrage** $\subseteq$ **tweet_id** $\times$ **entité** $\times$ {pertinent, non-pertinent}, pour les entités dont le nom peut porter à confusion, cette problématique se rapproche des tâches de désambiguïsation [10]. Il est important de ne pas ra-

8. Disponible publiquement à l'adresse suivante : http://nlp.uned.es/replab2013/ et http://nlp.uned.es/replab2014/

9. http://www.clef-initiative.eu/

10. Dans les cas de Apple ou encore Jaguar, il convient de s'assurer que l'on parle bien de la marque et non du fruit ou de l'animal.

ter l'information pertinente mais il est également important d'éliminer le maximum de bruit pour faciliter les traitements suivants ;

— **Opinion** $\subseteq$ **tweet_id** $\times$ **entité** $\times$ {Positive, Neutre, Négative}, ce problème est plus difficile à traiter que les problèmes usuels de détection d'opinion dans une critique de produit car même des nouvelles *à priori* factuelles peuvent se relever désastreuses pour la réputation d'une entité. La réciproque est également vraie lorsqu'un internaute s'attriste de la mort d'un artiste. Le message semble triste et négatif alors qu'il montre que l'utilisateur appréciait l'artiste ;

— **Regroupement** $\subseteq$ **tweet_id** $\times$ **tweet_id** $\times$ {discussion}, l'objectif est ici de regrouper les tweets similaires, par conversation, thème ou événement ;

— **Priorité** $\subseteq$ **tweet_id** $\times$ **cluster_id** $\times$ **entité** $\cup$ {Aucune, Moyenne, Élevé (ou Alerte)}, il s'agit ici de déterminer l'importance d'un message ou d'une discussion au regard de la réputation de l'entité. Il s'agit des messages traitant d'un sujet brûlant ou ceux pour lesquels une réaction immédiate est nécessaire, en clair si il y a des messages à ne pas laisser passer ce sont bien ceux-là. De plus, les organisateurs de RepLab indiquent qu'une opinion négative venant d'un internaute jugé important ou influent peut justifier le niveau de priorité élevé d'un message ;

— **Thématique** $\subseteq$ **tweet_id** $\times$ **entité** $\times$ { Citoyenneté, Gouvernance, Innovation, Leadership, Performance, Produits&Services, Satisfaction, Non Définie} Ces catégories correspondent aux points d'intérêts des investisseurs et actionnaires. Voir le bilan de (Amigó et al., 2014) pour plus de détails.

Près de 140, 000 tweets ont été manuellement annotés par les spécialistes du cabinet de e-Reputation espagnol Llorente & Cuenca [11] en fonction de chacun des critères correspondant aux problématiques énoncées ci-dessus. Les annotateurs étaient assistés par ORMA (Carrillo-de Albornoz et al., 2014), un système automatique d'annotation. La collection se divise en quatre parties, chacune correspondant à une des quatre thématiques suivantes : *automobile*, *banque*, *université* et *artistes*. Pour chacune des entités, les organisateurs de RepLab fournissent 2200 tweets écrits en langue anglaise ou espagnole (la répartition est d'environ 80-20% en faveur de l'anglais). Ces tweets ont été récupérés depuis Twitter à partir du *nom canonique* de l'entité entre le 1er juin et le 31 décembre de l'année 2012. Les 700 premiers messages constituent l'ensemble d'entraînement, les autres l'ensemble d'évaluation. Le tableau 3.4 montre la distribution des catégories thématiques dans les deux ensembles de données. L'annotation en catégorie thématique ne concerne les messages relatifs aux domaines automobile et bancaire. Notons la sur-représentation des « *Products & Services* » qui représentent presque 50% des données.

11. http://www.llorenteycuenca.com/

Catégorie	Description	Entraînement	Évaluation
Citoyenneté	Responsabilité sociétale	2,209	5,027
Gouvernance	Transparence et valeurs éthiques	1,303	3,395
Innovation	Adaptation au marché	313	1,943
Leadership	Gestion organisé et vision du future	297	763
Performances	Performances financières et rentabilité	947	1,598
Produits et SAV	Qualité des produits et du service clients	7,898	15,903
Satisfaction	Satisfaction et fierté des salariés	468	1,124
Non-Définie	-	2,268	4,349
Total	-	15,703	34,102

TABLE 3.4 – *Répartition des catégories thématiques dans les ensembles d'entraînement et d'évaluation.*

Le tableau 3.5 montre la distribution des opinions et des niveaux de priorité dans les ensembles d'entraînement et d'évaluation de la collection de données. A noter, le niveau de priorité élevé est la plus petite catégorie (environ 5% des données) concernant majoritairement les banques.

Priorité	**Entraînement**	**Évaluation**
Alerte	1,540	3,240
Important	17,961	38,617
Sans importance	15,379	33,613
Total	34,880	75,470
Opinion	**Entraînement**	**Évaluation**
Négative	5,409	11,006
Neutre	9,753	20,740
Positive	19,718	43,724
Total	34,880	75,470

TABLE 3.5 – *Répartition des opinions et des niveaux de priorité dans cette collection.*

La collection RepLab comporte 9,760 paquets (ou groupements de messages) pour les 134,000 messages de la collection (soit environ 18 messages par paquet). Une difficulté s'ajoute avec la bilingualité des données, car si les étiquettes sont exprimées en anglais, les paquets qu'elles désignent peuvent contenir des messages en espagnol et anglais ce qui perturbe totalement la plupart des algorithmes automatiques fonctionnant par similarité entre messages. Si l'on ramène le problème au niveau de chacune des 61 entités considérées dans la collection, nous avons un total de 160 paquets par entité pour les 2,200 messages relatifs à cette entité. Chaque message n'appartient qu'à un et seul paquet. Durant cette annotation, les annotateurs du cabinet Llorente & Cuenca ont, à l'aide d'outils automatiques, regroupé les messages relevant d'une même conversation d'un événement très spécifique par paquets de taille variable (de 1 à 150 messages) avant d'étiqueter le paquet à partir du message qui semblait être le plus représentatif du paquet. Une fois l'ensemble des paquets constitués, les annotateurs ont effectué une relecture visant à vérifier la cohérence des paquets. Durant cette phase de relecture, si un message apparaissait plus pertinent dans un autre groupe, ce message était alors affecté dans un autre groupe.

3.3.3 Collection RepLab'2014 Profils d'utilisateurs Twitter

Nous décrivons dans cette sous-section la partie de collection des données CLEF RepLab 2014 [12] (Amigó et al., 2014) dans le contexte de l'édition 2014 de *Conference and Labs of the Evaluation Forum* [13] (CLEF). Cette collection a été conçue pour une campagne d'évaluation dédiée au profilage d'utilisateurs . Le problème a été découpé par les organisateurs de RepLab en deux sous problématiques : la détection d'influence et la catégorisation socio-professionnelle des utilisateurs. Ces deux problématiques ont pour objectif d'amener les chercheurs à extraire des caractéristiques indiquant le niveau d'influence réelle d'utilisateurs et/ou la catégorie socio-professionnelle d'utilisateurs du réseau social Twitter

Près de 7,000 utilisateurs différents, actifs dans les domaines de l'automobile et de la banque [14], ont été manuellement annotés par les spécialistes du cabinet de e-Reputation espagnol Llorente & Cuenca [15] en fonction d'une catégorie socio-professionnelle (parmi les 9 évoquées ci-dessous) et de l'influence selon laquelle ils sont censés être perçus et non pas selon l'influence qui pourrait être mesurée à partir des données de leurs comptes Twitter [16]. Les organisateurs de RepLab donnent également accès aux 600 derniers tweets de chaque utilisateur au moment où ils ont constitué la collection. Ces tweets peuvent avoir été écrits en anglais ou en espagnol.

La collection se divise en deux parties, chacune correspondant à une des deux thématiques évoquées plus haut, automobile et banque. Cette classification a été effectuée manuellement en fonction de la thématique d'activité principale de l'émetteur des tweets. Ces thématiques sont exclusives, c'est à dire qu'un pseudo n'appartient qu'à une et une seule thématique. Les deux sous-collections sont relativement équivalentes en termes de taille, avec de chaque côté 1,186 et 1,314 utilisateurs pour les 2,500 constituant le jeu d'entraînement et 2,353 et 2,507 respectivement pour automobile et banque pour les 4,500 utilisateurs du jeu d'évaluation. Les utilisateurs de l'ensemble d'évaluation étant, bien entendu, différents de ceux de l'ensemble d'entraînement. La collection comporte au final près de 4,000,000 de messages ce qui n'est pas sans conséquences sur les temps de traitement de certaines approches.

12. Disponible publiquement à l'adresse suivante : http://nlp.uned.es/replab2014/
13. http://www.clef-initiative.eu/
14. Les profils sélectionnés ont au moins 1,000 abonnés.
15. http://www.llorenteycuenca.com/
16. Ni même l'influence établie avec des outils (tels que Klout et Kred) que nous avons évoqués précédemment.

Si le challenge est bien une question de classement par niveau d'influence, l'annotation pour la tâche de détection d'influence est binaire. Un utilisateur est «*influant*» ou «*non-influant*»[17]. Cette annotation est disponible pour un jeu d'entraînement dont 796 des 2500 utilisateurs sont considérés comme «*influents*». L'annotation est masquée pour les 4500 utilisateurs restants (dont 1563 «*influents*») considérés comme le jeu d'évaluation.

Dans le chapitre 6, nous nous limiterons aux 4,720 utilisateurs (respectivement 2,310 pour le domaine automobile et 2,410 pour les banques) pour lesquels nous avons pu collecter l'ensemble des informations (caractéristiques de profil et estimation du niveau d'influence par des méthodes automatiques basées sur l'analyse de contenu détaillées dans les chapitres 2 et 5). Nous ne nous intéresserons alors plus qu'à une frange plus restreinte des messages de chaque utilisateur et nous vérifierons la stabilité des performances des méthodes proposées.

3.4 Sources d'informations additionnelles et contextualisation

Dans le cheminement qui amène l'annotateur (ou l'analyste) à prendre une décision, nous avons vu que le contenu textuel[18] tient une place primordiale toutefois il arrive que cela ne soit pas suffisant. L'analyste fait appel à sa mémoire pour remettre en contexte le message, mais il arrive également que l'analyste cherche des informations supplémentaires pour compléter le contenu et aider à sa compréhension.

3.4.1 Contextualisation de Micro-Blogs : le cas INEX Tweet Contextualization 2014

Comprendre comment l'analyste résout les problèmes de classification de contenus liés à l'image de marque et proposer des méthodes permettant d'approcher ce raisonnement était un des objectifs principaux de RepLab (Amigó et al., 2013a). Toutefois, à l'instar de méthodes complexes et à la différence d'un

17. Notons que l'influence fournie est binaire, un profil est noté comme étant «*influant*» ou «*non-influant*» alors qu'il est demandé d'ordonner les utilisateurs selon leur niveau d'influence. Notons également, que RepLab 2014 s'est déroulé en partenariat avec PAN 2014. le bilan commun à l'issue des deux campagnes a fait émerger la corrélation entre l'influence d'une personne dans les domaines économiques avec son âge et son sexe : les utilisateurs influents dans ces milieux sont des hommes de moins de 45 ans (Rangel et al., 2014).

18. Accompagné de ses méta-données immédiates (l'auteur, la date).

simple classifieur tels que ceux que nous avons vus au cours du chapitre 2, l'analyste a recours à de l'information contextuelle. Information qui fait partie de son savoir faire (expérience) ou qu'il peut aller chercher volontairement pour compléter le contenu. L'objectif du challenge « *INEX Tweet Contextualization* 2014 » (Bellot et al., 2014) était de répondre à la question : « *De quoi parle ce tweet ?* » c'est à dire contextualiser automatiquement des messages publiés sur Twitter. Pour un message m, cela revient à fournir de l'information contextuelle afin d'aider le lecteur à mieux comprendre les tenants et les aboutissants du sujet principal dont il est question dans le message. Dans le cadre du challenge, ce contexte devait prendre la forme d'un résumé lisible de quelques centaines de mots extraits à partir de Wikipédia. Le résumé étant spécifique à chaque message, il doit donc comporter un maximum d'informations pertinentes au regard de ce dernier.

Collection de données

La collection de données utilisée durant le challenge est composée des deux parties décrites ci-dessous :

- Un extrait du Wikipédia en anglais datant de novembre 2012 (nettoyé des pages vides, notes et référence bibliographiques) ;
- Une sélection de 240 messages issus de la collection RepLab 2013, ces messages comportent au minimum 80 caractères et ne contiennent aucun lien.

L'entité à laquelle le message est rattaché sert de point d'entrée sur Wikipédia. C'est à partir de ce point d'entrée que les systèmes automatiques devront produire une perspective contextuelle. Les résumés contextuels ont été officiellement évalués durant le challenge selon les critères suivants :

- Informativité : quantité d'informations pertinentes qu'ils contiennent
- Lisibilité selon plusieurs métriques :
  - Syntaxe : présence ou non de problèmes de syntaxe ;
  - Anaphore : la présence d'anaphores non résolues ;
  - Redondance : la quantité d'informations redondantes ;
  - « *Trash* » : le passage sélectionné n'a aucun lien avec le message qu'il doit contextualiser

3.4.2 Contextualisation de Micro-Blogs : la généralisation lexicale

Comme nous l'avons vu en section 3.3.2, l'ensemble d'évaluation de RepLab est deux fois plus riche que l'ensemble annoté d'entraînement à notre disposi-

tion. Nous pouvons supposer qu'il existe la même relation concernant la diversité du vocabulaire. Cette différence engendre un phénomène de mots hors vocabulaire bien connu (plus commun dans la littérature sous la dénomination « *Out Of Vocabulary words* » ou « *OOV* ») qui implique une perte d'information. En complément des expériences précédentes, nous proposons donc d'utiliser le contexte lexical des messages pour réduire cette perte d'information.

Chaque mot hors vocabulaire connu est projeté dans l'espace du vocabulaire connu à travers le modèle de représentation continue des mots décrit par (Bengio et al., 2003) en considérant ce modèle comme un moteur de généralisation de mots. Nous utilisons le modèle Word2Vec (Mikolov et al., 2013b) qui est appris à partir d'un réseau de neurones basés sur des skip-gram. Ce réseau de neurones essaye de maximiser la probabilité suivante (Mikolov et al., 2013b) :

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{-c<j<c, j\neq 0} log(\frac{exp(i_{w_{t+j}}^{T} o_{w_t})}{\sum_{w=1}^{N} exp(i_{w}^{T} o_{w_t})}) \tag{3.9}$$

Où N est le nombre de mots présents dans l'ensemble d'entraînement, $w_0..w_N$ une séquence de mots dans cet ensemble et c la taille du contexte lexical considéré. (Mikolov et al., 2013b) ont déjà montré la capacité des modèles Word2vec à capturer les relations syntaxiques et sémantique entres les mots. Ces modèles nous permettent d'évaluer la similarité sémantique des mots à partir d'opérations géométriques simples comme les mesures d'angles.

Chapitre 4

Catégorisation de Micro-Blogs, un problème de messages ?

Sommaire

Nous présentons dans cette section les différentes expériences que nous avons menées afin d'évaluer l'efficacité des méthodes automatiques de catégorisation dans le but de valider notre hypothèse : être en mesure de prédire automatiquement l'opinion exprimée dans un contenu textuel à partir de l'analyse statistique des termes qui le composent. Nous utilisons les méthodes décrites dans le chapitre 2 et les protocoles d'évaluation (sources d'informations et mesures de performances) vus dans le chapitre précédent afin de rester dans le cadre d'usage de la littérature. Nous commençons par proposer quelques analyses reposant sur les expériences de catégorisation thématique sur les collections Imagiweb et RepLab. Nous proposons par la suite d'analyser les performances de ces mêmes systèmes confrontés à une tâche d'analyse de sentiment sur des contenus de type Micro-Blogs réputés difficiles à manipuler. Nous proposons ensuite des analyses complémentaires en ayant recours à des méthodes d'enrichissement de ces contenus à partir des ressources que nous venons de présenter dans la section précédente.

4.1 Evaluation

4.1.1 Performances de catégorisation thématique

Nous nous intéressons dans cette section aux données Imagiweb et RepLab, ces dernières présentant un état des lieux des récents progrès reportés en la matière dans la littérature. Les performances obtenues par nos méthodes sur les données RepLab lors de la campagne sont également reportées dans (Cossu et al., 2014) puis hors campagne dans (Cossu et al., 2015a).

Cas RepLab

Dans la tâche de détection thématique de RepLab 2014, nous rencontrons deux problèmes intéressants. Tout d'abord, lors du concours, les organisateurs ont décidé d'exclure la catégorie thématique « *Non Définie* » de l'évaluation officielle afin de de n'évaluer « *l'Accuracy* » [1] des systèmes que sur les 7 catégories thématiques restantes. D'autre part, cette précision est considérée au niveau globale, c'est à dire qu'elle correspond à la précision moyenne des précisions obtenues sur les données de entité prise individuellement.

Malgré cette évolution faite à l'échelle de l'entité, et étant donné qu'il n'y a pas assez de données permettant d'entraîner une approche automatique, nous

1. Notée « *Précision* » dans le tableau 4.1.

avons considéré dans nos approches que le vocabulaire propre aux thématiques ne dépend pas d'une entité spécifique mais que ces thématiques sont exprimées globalement.

Nous avons ensuite proposé de compléter l'évaluation officielle avec un Macro F-Score (noté F-Score) permettant de comparer la capacité des systèmes à récupérer l'information sur l'ensemble des 7 catégories considérées [2]. Nous proposons dans un premier temps de tirer parti de cette évaluation et de considérer le cas de la catégorie « *Non Définie* » comme relevant d'un problème de filtrage, c'est à dire que :

— (i) les données d'entraînement concernant cette catégorie sont ignorées, au message ;
— (ii) tous les messages de la collection d'évaluation sont catégorisés comme appartenant l'une des catégories ;
— (iii) les messages qui ne sont pas susceptibles de pouvoir entrer dans une des catégories retenues sont ignorés par le système.

Dans un second temps, nous proposons, plutôt que d'ignorer cette catégorie, de la considérer comme faisant partie du problème de catégorisation. C'est à dire que nous évaluons également la capacité des systèmes automatiques à distinguer cette catégorie parmi les autres. Il serait également possible d'imaginer un processus itératif d'apprentissage actif dans lequel des analystes feraient émerger de nouvelles catégories à partir de ce que les systèmes considèrent comme « *non défini* ».

Ce choix de double évaluation permet finalement d'observer des comportements intéressants. La première série d'expériences (évaluations notées (-U) dans le tableau 4.1) ne s'est pas vraiment révélée concluante car, malgré la complexité réduite (espace des modèles à 7 dimensions au lieu de 8) par rapport à la seconde série (évaluations notées (+U)), les performances ne sont pas significativement meilleures. Même si l'on observe un léger mieux dans le premier cas.

Nous comparons également les performances de ces systèmes avec les résultats obtenus par les meilleurs participants à la campagne : (McDonald et al., 2015; Rahimi et al., 2014) avec leurs systèmes d'enrichissement de messages mais également avec les baselines proposées par les organiseurs [3]. A noter, les différences entre les performances des approches cosine et SVM ne sont pas significatives.

2. Ce F-Score est calculé au niveau des catégories thématiques sans distinction de l'entité à laquelle le message est rattaché.

3. La baseline naïve consiste à affecter tous les messages à la catégorie majoritaire. La baseline SVM revient à entraîner un SVM linéaire pour chaque entité en utilisant la présence des mots (valeur binaire) (cf. (Amigó et al., 2014)).

Approche	F-Score (-U)	Précision (-U)	F-Score (+U)	Précision (+U)
cosine	**.491**	**.736**	**.500**	**.693**
SVM	.469	.732	.461	.679
(McDonald et al., 2015)	.473	.731	-	-
(Rahimi et al., 2014)	.489	.695	-	-
SVM	.38	.622	-	-
Naive	.152	.560	-	-

TABLE 4.1 – *Performances des systèmes automatique de catégorisation thématiques sur les données d'évaluation de la collection RepLab* 2014 *ordonnés par Précision (-U). Les méthodes utilisant la généralisation lexicale sont notées (w/ Context). Les scores les plus élevés sont indiqués en gras.*

Cas Imagiweb Twitter

Contrairement au cas précédent, les données Imagiweb permettent de constituer des modèles spécifiques à chaque entité. A titre indicatif, en nous basant sur la même supposition d'indépendance des thématiques *vis à vis* des entités nous avons effectué des expériences (i) en inversant puis (ii) combinant les modèles appris pour chaque entité. C'est à dire dans le premier cas : retrouver la thématique dans un message concernant Nicolas Sarkozy en utilisant les données de François Hollande voire dans le second cas les données des deux entités à la fois. Cette situation nous replace cadre de RepLab où lors de nos expériences, nous avions considéré les données de l'ensemble des entités.

Ce scénario revient à considérer que ce qui est utilisé pour exprimer les thématiques concernant une entité, permet d'enrichir la sémantique de ces mêmes thématiques exprimées *vis à vis* d'autres entités. Finalement, sur les données Imagiweb, les modèles dits « *combinés* » ont systématiquement montré une légère amélioration des performances de catégorisation et ce pour les deux entités. De ce fait, nous ne présentons dans le tableau 4.2 que des résultats de nos systèmes de catégorisation thématique basés sur ces modèles. Les résultats présentés concernent les deux entités (François Hollande et Nicolas Sarkozy) étudiées sur les ensembles de développement et évaluation.

Développement	François Hollande		Nicolas Sarkozy	
Approche	**F-Score**	**Précision**	**F-Score**	**Précision**
cosine	.259	.325	.299	.430
K-PPV	.288	.374	.310	.458
Fusion	.290	.391	.328	.448
Evaluation	François Hollande		Nicolas Sarkozy	
Approche	**F-Score**	**Précision**	**F-Score**	**Précision**
cosine	.367	.468	.280	.463
K-PPV	.343	.446	.289	.467
Fusion	.369	.473	.269	.451

TABLE 4.2 – *Performances des systèmes automatiques de catégorisation thématique utilisant des modèles combinés sur les sous-collections de développement et d'évaluation de la collection Imagiweb Twitter. Les scores les plus élevés sont indiqués en gras. Les différences de performances entre les systèmes présentés ne sont pas significatives.*

La combinaison de systèmes montre toute sa robustesse en complément d'une amélioration significative des performances de catégorisation. Si l'on s'intéresse à nos scénarios de validation, (dans un premier cas l'expert chargé de la vérification des données avait accès à l'hypothèse automatique alors qu'il annotait à l'aveugle dans le second cas) les différences sont significatives pour la catégorisation thématique où le Macro F_Score varie entre .255 pour l'annotation

à l'aveugle et .334 lorsque ce dernier avait accès l'hypothèse générée automatiquement. Ce résultat reste toutefois à mettre en relation avec la difficulté de la tâche, les thématiques étant vagues et un message relevant souvent de plus d'une seule thématique. Nous pouvons considérer que l'annotateur a simplement validé l'hypothèse proposée par facilité car celle-ci n'était pas forcément fausse mais que d'autres pouvaient également correspondre.

Cas Imagiweb Blogs

Nous présentons dans le tableau 4.3 les résultats de nos systèmes de catégorisation thématiques appliqués sur les données EDF. Nous avions vu en section 3.3.1 que cette collection de données a été annotée selon deux niveaux de thématiques (thématiques et sous-thématiques). Il est apparu durant la phase de développement que détecter les trois thématiques (Risques, Stratégies et Tarifs) était une tâche relativement triviale. Nous avons donc décidé de nous concentrer sur la détection de sous-thématiques (en tenant compte ou non de l'information de thématique « *parente* »). Sans tenir compte de cette information, nous avons remarqué des affectations atypiques résiduelles c'est à dire que certaines sous-thématiques ont été rattachées par les classifieurs à une thématique parente différente de celle définie dans la grille d'annotation. Par exemple l'association des tarifs et des déchets permet de mettre en évidence le lien argumentatif fait par certains blogueurs entre la question des coûts de production et de la gestion des déchets, au-delà de la seule problématique des risques. Ces découplages nous permettent une analyse intéressante de cette grille d'annotation car cela révèle les croisements thématiques qui sont parfois opérés par les auteurs des messages. Si l'on s'intéresse aux scénarios de validation (Aveugle ou non), les différences sont observées sur un trop petit nombre de documents pour être significatives.

4.1.2 Performances de détection d'opinions

Nous ne nous intéressons dans cette section qu'à l'évaluation de nos méthodes appliquées aux données Imagiweb. Les performances obtenues par nos méthodes sur les données RepLab lors de la campagne (cf. (Cossu et al., 2013)) ont été rendues obsolètes par les récentes avancées proposées par la littérature (Gârbacea et al., 2014). Ces dernières considèrent depuis la question comme étant résolue (Spina, 2014; Peetz, 2015).

Développement		
Approche	**Précision**	**F-Score**
cosine	.574	.541
Fusion	**.591**	**.590**
K-PPV	.537	.524
Evaluation		
Approche	**Précision**	**F-Score**
cosine	.627	.606
Fusion	**.716**	**.701**
K-PPV	.580	.572
Fusion		
Aveugle	.651	.592
Validation	.746	.708

TABLE 4.3 – *Performances des systèmes automatiques de catégorisation thématique sur les sous-collections de développement et d'évaluation de la collection Imagiweb EDF. Les scores les plus élevés sont indiqués en gras. Les différences de performances entre les systèmes présentés ne sont pas significatives*

Cas Imagiweb Twitter

Nous rapportons les performances de nos systèmes automatique de catégorisations d'opinions (polarité) dans le tableau 4.4. Les résultats présentés concernent les deux entités étudiées sur les ensembles de développement et évaluation.

Développement	François Hollande		Nicolas Sarkozy	
Approche	**F-Score**	**Précision**	**F-Score**	**Précision**
cosine	.546	.664	.562	.618
Fusion	**.547**	**.672**	**.570**	**.625**
K-PPV	.534	.662	.528	.590
Evaluation	François Hollande		Nicolas Sarkozy	
cosine	**.535**	.754	.504	.617
Fusion	**.535**	**.757**	**.520**	**.620**
K-PPV	.495	.758	.490	.607

TABLE 4.4 – *Performances des systèmes automatiques de détection d'opinion utilisant des modèles spécifiques sur les sous-collections de développement et d'évaluation de la collection Imagiweb Twitter. Les scores les plus élevés sont indiqués en gras. Les différences de performances entre les systèmes présentés ne sont pas significatives.*

Comme pour la catégorisation thématique, nous avons essayé différentes stratégies concernant les modèles d'apprentissage à commencer par utiliser les

modèles de l'autre entité pour effectuer la détection d'opinion. L'objectif était ici de mesurer les performances des systèmes en utilisant des données inadaptées (vocabulaire complètement différent et opinions opposées sur des thématiques comme les projets et bilans économiques). Logiquement, les résultats observés sont largement inférieurs. Cela était toutefois attendu, les messages portant sur des thèmes tels que les bilans et projets politiques amènent des considérations diamétralement opposées selon le candidat : ce qui est positif pour l'un et généralement négatif pour l'autre alors que c'est pourtant le même vocabulaire qui est utilisé.

Si l'on s'intéresse de nouveau aux scénarios de validation (hypothèse connue ou masquée), les différences ne sont pas significatives pour la détection de polarité (précision entre .62 et .634 pour la combinaison de méthodes automatiques). A titre informatif, nous pouvons là encore noter que dans le cas d'une prédiction conjointe thématique et polarité, la combinaison linéaire obtient au mieux (pour Nicolas Sarkozy) un macro F-Score moyen de .17 et une précision de .35. La marge de progression à ce sujet reste donc très grande.

Cas Imagiweb EDF

Nous présentons dans le tableau 4.5 les résultats de nos systèmes de catégorisation d'opinions appliqués sur les données EDF. Les blogs étant des documents assez long (plusieurs phrases portant souvent chacune sur un fait différent) l'analyse de polarité (donc contenu subjectif) est une tâche plus difficile que celle visant à déterminer la thématique (une idée générale, souvent assez objective). Notons à titre indicatif que dans le cas d'une prédiction conjointe thématique et polarité, la combinaison linéaire (notée Fusion dans les tableaux 4.5 et 4.3) obtient un macro F-Score moyen de .49 et une précision de .68. Ces chiffres indiquent une certaine robustesse de nos méthodes à trouver conjointement les bonnes catégories d'opinions.

4.1.3 Performances de filtrage et détection d'alerte (ou priorité)

Nous pouvons considérer ces deux problématiques comme étant très similaires, la seconde n'étant qu'un approfondissement nuancé de la première. Notons toutefois que contrairement au filtrage qui est un des problèmes les plus abordés en RI (cf. campagnes TREC et KBA), la question de la détection d'alerte reste encore assez peu explorée du fait de sa complexité. En effet, distinguer deux messages entre eux par le niveau d'importance que l'on doit respectivement leur accorder est un challenge flou et dépendant des considérations de

Développement		
Approche	**Précision**	**F-Score**
cosine	.804	.683
K-PPV	.777	.621
Fusion	**.834**	**.732**
Evaluation		
Approche	**Précision**	**F-Score**
cosine	.804	.683
K-PPV	.783	.633
Fusion	**.865**	**.763**
Fusion		
Aveugle	.781	.627
Validation	.852	.796

TABLE 4.5 – *Performances des systèmes automatiques de catégorisation thématique sur les sous-collections de développement et d'évaluation de la collection Imagiweb EDF. Les scores les plus élevés sont indiqués en gras. Les différences de performances entre les systèmes présentés ne sont pas significatives.*

chacun. Dans le cadre de RepLab, cette notion d'importance est même spécifique à chaque entité alors que l'on aurait pu intuitivement penser que certaines thématiques sont importantes quelque soit l'entité d'un même domaine (cf. section 6.6.3).

Nous reportons dans le tableau A.2 les performances des méthodes que nous avons présentées sur la tâche de filtrage de RepLab 2013. Notons que la répartition entre documents pertinents et non-pertinents est de 78-22%. L'ensemble des méthodes proposées obtiennent des performances supérieures à celles de la médiane des participants. Une seule de nos approches (celle basée sur les plus proches voisins) n'arrive toutefois pas à dépasser la fonction témoin proposée par les organisateurs (étiquette du document le plus proche dans l'ensemble d'entraînement). Nous avons également proposé des combinaisons plus avancées de ces approches que nous décrivons dans (Cossu et al., 2015a). Comme l'ont montré (Qureshi et al., 2014), logiquement les méthodes, comme celles de (Hangya et Farkas) et (Saleiro et al., 2013), utilisant des ressources additionnelles spécifiques [4] à chaque entité obtiennent de meilleures performances. D'ailleurs la littérature considère qu'avec ce niveau de résultat, dépassant les taux d'accords entre deux annotateurs constatés sur le problème, le problème du filtrage de Micro-Blogs est résolu (Spina, 2014).

Durant l'édition 2013 de RepLab, la détection de priorité a été abordée comme un problème de catégorisation de message par la plupart des parti-

4. En complément du contenu textuels des messages.

Approche	Précision	F(R,S)
(Saleiro et al., 2013)	.908	**.488**
(Hangya et Farkas)	**.928**	.438
(Cossu et al., 2013) *Lia_Filter_1*	.872	.381
Fonction témoin	*.871*	*.325*
(Cossu et al., 2013) *Fusion PROMETHEE*	.882	.312
(Cossu et al., 2013) *Fusion ELECTRE*	.879	.302
(Cossu et al., 2013) *Combinaison Linéaire*	.874	.296
(Cossu et al., 2013) *KBA*	.850	.289
cosine	.835	.272
Mediane	*.826*	*.265*

TABLE 4.6 – *Performances des systèmes de filtrage sur les données d'évaluation de la collection RepLab* 2013 *ordonnées selon la métrique officielle du challenge : F(R,S). Les meilleurs performance sont indiquées en gras. Les différences de performances entre les systèmes présentés ne sont pas significatives.*

cipants. Nous reportons dans le tableau 4.7 les résultats obtenus par les méthodes que nous proposons, nous comparons ces résultats avec ceux obtenus durant la campagne, notons que toutes nos propositions obtiennent des résultats supérieurs au niveau médian des résultats des participants. L'équipe LIA (Cossu et al., 2013) avec une approche de catégorisation basée sur les k plus proches voisins (cette dernière est notée Lia_Prio_5 dans le tableau 4.7) a obtenu la meilleure performance reconnue à ce jour selon la métrique officielle *F(R,S)*. Si l'on s'intéresse aux performances des autres méthodes reportées dans le tableau 4.7 nous constatons que les approches basées sur des SVM ou la distance cosine obtiennent des résultats tout à fait comparables voire meilleurs en précision. Toutefois selon le *F-Score* ou la mesure officielle, ces dernières restent moins performantes, cosine étant même moins performante que la fonction témoin proposée par les organisateurs en termes de métrique *F(R,S)*. Comme pour la détection thématique vue en section 4.1.1, nous reportons également précision moyenne des précisions par entité. Il est intéressant de noter que les combinaisons proposées par (Cossu et al., 2013) obtiennent des scores plus élevés en précision mais reste largement en dessous en termes de performance *F(R,S)*. Notons enfin la contre-performance de l'approche « *KBA* » en terme de F-Score et de précision, alors que la méthode apparaît pourtant comme compétitive selon la mesure *F(R,S)*.

Il est également possible de considérer cette détection selon plusieurs angles, en privilégiant l'un ou l'autre des critères de rappel et précision. Maximiser la précision au détriment du rappel implique de ne présenter à l'utilisateur final qu'uniquement un petit nombre de messages dont nous pouvons assurer la qualification « *d'alertes* », quitte à rater certains contenus. A l'inverse, maximi-

Approche	F-Score	Précision	F(R,S)
(Cossu et al., 2013) *Lia_Prio_5*	**.571**	.636	**.335**
SVM	.563	.644	.304
(Cossu et al., 2013) *KBA*	.421	.585	.282
Fonction témoin	.512	.570	.274
cosine	.561	.633	.260
(Cossu et al., 2013) *Fusion PROMETHEE*	-	.651	.253
(Cossu et al., 2013) *Combinaison Linéaire*	-	.647	.251
(Cossu et al., 2013) *Fusion ELECTRE*	-	**.652**	.251
Mediane	-	*.573*	*.249*

TABLE 4.7 – *Performances des systèmes de détection de priorité sur les données d'évaluation de la collection RepLab* 2013 *ordonnées selon la métrique officielle du challenge : F(R,S). Les meilleurs performance sont indiquées en gras. Les différences de performances entre les systèmes présentés ne sont pas significatives.*

ser le rappel consiste à présenter beaucoup plus de messages pour être sur de ne rien rater mais en prenant le risque de faire perdre son temps à l'expert en lui suggérant des messages totalement inintéressants. Le nombre relativement restreint de messages étiquetés comme étant des *alertes* permet également de considérer le problème comme étant une recherche d'information et potentiellement de proposer une évaluation de pertinence des messages renvoyés.

Sur d'autres collections de données également extraites du réseau Twitter, (Weerkamp et al., 2011) ont observé des comportements distincts des différents utilisateurs de Twitter en fonction de leur nationalité (déduite de la langue dans laquelle sont écrits la majorité de leurs tweets). Leurs expériences les ont amenés à formuler les conclusions suivantes : les messages écrits en langue allemande sont généralement longs et bien structurés à l'instar de conversation entre professionnels. Ils sont de plus caractérisés par l'utilisation fréquente de mots-dièses et liens. A l'inverse en Espagnol, les messages sont écrits sans véritable structure. Il s'agit plutôt de conversations personnelles avec notamment beaucoup de mentions entre interlocuteurs. Dans notre cas, à partir de nos observations sur la collections RepLab'2013 nous pouvons noter que la plupart des alertes sont des messages longs en espagnol. Cette information supplémentaire pourrait être prise en considération par des méthodes comme celles décrites en section 2.2.5 et puis intégrées dans la combinaison que nous proposons pour affiner l'analyse faite au niveau du contenu.

Nous avons montré par nos expériences que les méthodes proposées, qui d'une relative faible complexité par rapport à celles que l'on peut trouver dans la littérature, offrent des niveaux de performances similaires si ce n'est parfois meilleurs que ces dernières. Cette observation se justifie sans doute par le peu de données disponibles pour entraîner de manière plus robuste les méthodes

les plus complexes.

4.2 Message ou discussion quelle granularité pour la détection de priorité ?

Nous nous intéressons ici de nouveau à la question de la détection du niveau de priorité d'un message telle que définie dans le cadre de la tâche du même nom du challenge RepLab 2013. Nous avons vu en section 4.1.3 qu'il est possible d'estimer le niveau de priorité d'un message à partir de son contenu textuel. Toutefois, ce dernier est souvent insuffisant pour justifier correctement le niveau de priorité élevé dont relève un message. C'est pourquoi, nous envisageons maintenant d'aborder cette question à partir d'un niveau de granularité différent. Nous passons de l'échelle d'un seul contenu à celle d'un groupe de plusieurs messages.

Afin d'étudier la dépendance existante entre l'attribution d'un niveau de priorité et les regroupements thématiques, nous proposons plusieurs méthodes automatiques supervisées ou non. Nous considérons également la fonction témoin proposée par les organisateurs de la tâche ; cette dernière consiste à rattacher chaque message de l'ensemble d'évaluation à celui qui lui ressemble le plus dans l'ensemble d'entraînement (selon l'indice de similarité Jaccard).

4.2.1 Méthodes automatiques de regroupement de messages

Nous proposons de regrouper les messages en utilisant tout d'abord un algorithme de regroupement utilisant la distance Jaccard comme mesure de similarité entre les messages. Contrairement à la fonction témoin proposée par les organisateurs, nous proposons de calculer cette distance à partir de la représentation vectorielle discriminante telle que nous l'avons présentée dans le chapitre 2. Cette méthode est appliquée sur l'ensemble des données de la collection (entraînement et évaluation). La valeur k du nombre de groupes à retrouver est fixée au nombre de groupes présents dans l'ensemble d'entraînement.

La deuxième méthode que nous proposons est un algorithme de classification hiérarchique basé sur la même distance. Cette fois-ci, l'algorithme regroupe les messages par paires puis paires de paires jusqu'à obtenir les k groupes souhaités.

Nous proposons une troisième méthode basée sur un mécanisme de sélection de termes appelé **Maximum *a posteriori* (notée MAP X par la suite)**. Elle

est utilisée pour la détection de thèmes en traitement automatique de la parole (Hazen et al., 2007). Contrairement aux deux méthodes précédentes, celle-ci nécessite une phase d'entraînement qui consiste à sélectionner les termes les plus représentatifs de chaque groupe à partir de leur probabilité d'appartenance aux différents groupes dans l'ensemble d'apprentissage. Les messages de l'ensemble d'évaluation sont donc assignés au groupe de la collection d'entraînement dont ils sont le plus proche.

4.2.2 Tâche de détection d'alertes

Nous utilisons les données et les mesures d'évaluation relatives à la tâche de détection d'alertes du challenge RepLab 2013 décrites dans le chapitre 3. Nous comparons la méthode que nous avons décrite précédemment avec la soumission officielle qui a obtenu les meilleurs résultats lors du challenge : l'approche «*LIA _Prio_5*» (Cossu et al., 2013) basée sur un algorithme de k plus proches voisins. Pour rappel, ce dernier consiste à rapprocher chaque message de l'ensemble d'évaluation des k plus proches messages de l'ensemble d'entraînement, ces derniers votant selon leur similarité pour le niveau de priorité duquel ils relèvent. A noter que la similarité est estimée selon l'indice de Jaccard à partir de la représentation discriminante du sac de mots du message auquel ont été ajoutés des jetons indicateurs de l'auteur du message et de l'entité à laquelle le message est rattaché. Pour cette expérience, nous avions fixé la valeur de k à 6 après validation sur l'ensemble de développement que nous avions établi à cet effet.

4.2.3 Regroupement de messages

Les données que manipulations comme «*matière brute*» sont les sorties des systèmes automatiques de détection de niveau de priorité. Nous souhaitons évaluer le gain de performances dans la tâche de détection de priorité en considérant l'information additionnelle apportée par les systèmes de groupement automatique de messages. Les performances de ces derniers sont reportées dans le tableau 4.8 y compris la fonction témoin mise au point par les organisateurs et le meilleur système officiellement soumis au challenge (Spina et al., 2013).

L'information additionnelle du groupe de messages est intégrée dans une combinaison qui consiste à trouver le niveau de priorité majoritaire à l'intérieur du paquet de messages et à le propager à l'ensemble des messages du groupe. Nous avons considéré que chaque message dans le groupe avait le même poids, d'autres solutions auraient pu être envisagées à ce niveau-là en utilisant par

exemple la probabilité de priorité fournie par les systèmes de détection automatique de priorité ou d'autres mesures permettant de définir la représentativité du message dans son paquet. Enfin, étant donné que les annotateurs ont considéré «*l'alerte*» comme étant une notion liée au sujet de conversation et non à un message spécifique nous aurions pu envisager qu'il suffit d'un message dont le niveau d'alerte a été estimé avec certitude pour décider que tout le groupe est tout aussi important que le message. Cette dernière méthode présente cependant l'inconvénient de propager des fausses alertes et donc de submerger l'utilisateur final (ici l'analyste de réputation) par des messages inutiles.

4.2.4 Évaluation des méthodes de regroupement de messages

Le premier point intéressant est que les différentes méthodes que nous proposons dépassent la performance obtenue par la fonction témoin et se positionnent à différents niveaux de Reliability (R) et Sensitivity (S). Nous pouvons d'ailleurs noter qu'il existe deux seuils différents concernant le système *MAP X* (variantes obtenues en jouant sur le nombre de mots retenus pour chaque catégorie dans l'ensemble d'entraînement). Dans le premier cas, c'est la Sensitivity qui est maximisée pour la variante MAP#1 ; dans l'autre, il s'agit de la Reliability (MAP#2) [5]. La contre performance des systèmes de classification non-supervisés s'explique la grande différence entre le nombre de paquets de messages entre les sous-collection d'entraînement et d'évaluation (du simple au double).

Approche	Reliability	Sensitivity	F(R,S)
Best@Replab2013	**.462**	.324	**.325**
MAP #1	.193	**.497**	.266
MAP #2	.381	.172	.238
Hierarchic clustering	.261	.220	.227
K-means clustering	.308	.157	.201
Replab baseline	.152	.217	.173

TABLE 4.8 – *Performances de nos systèmes automatiques de regroupement de messages comparés avec la baseline et le meilleur participant au challenge RepLab* 2013*. Les meilleurs scores sont indiqués en gras, les méthodes sont triées selon la mesure officielle du challenge : F(R,S). Les différences de performances entre les systèmes présentés ne sont pas significatives.*

Nous avons à titre expérimental tester d'autres configurations du système *MAP X*. Nous reportons sur la figure 4.1 les performances de ce système selon la mesure F(R,S) en fonction du nombre de termes considérés pour quali-

5. Ces deux variantes se distinguent par le nombre de termes retenus pour qualifier chaque classe très faible dans un cas (cinq), plus important dans l'autre (une cinquantaine).

fier chaque classe. Nous pouvons noté que nous arrivons très vite à une situation dite de «*sur-apprentissage*» car en considérant plus de termes, les performances augmentent sur la collection de développement sans que cela ne soit le cas pour la collection d'évaluation. Logiquement, considérer plus de données (termes) dans le modèle de qualification de chaque classe augmente la complexité de la méthode et donc les temps de traitement.

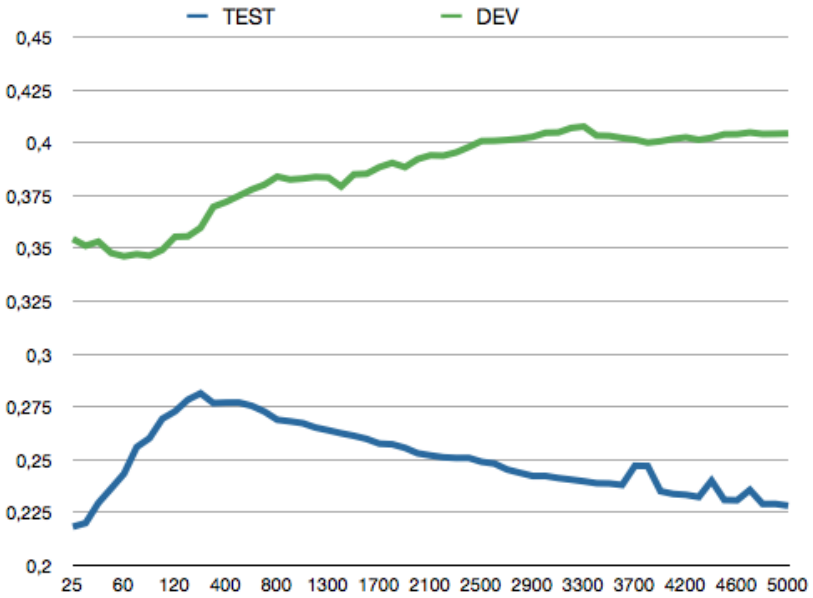


FIGURE 4.1 – *Evolution des performances (F(R,S)) sur les collections de développement et d'évaluation en fonction du nombre de termes utilisés pour qualifier chaque classe.*

4.2.5 Évaluation : la détection du niveau de priorité de groupes des messages

Nous souhaitons vérifier les performances d'une approche état de l'art, sur un type de documents totalement différent et sans véritable adaptation, comme l'approche décrite en section 2.2.5 proposée pour le challenge KBA [6]. Par rapport au challenge KBA où un système spécifique avait été entraîné pour chaque entité, nous avons pris le parti de considérer un système multi-entité afin de pouvoir repérer des marqueurs de pertinence indépendamment d'une entité spécifique et ainsi pouvoir travailler sur des entités pour lesquelles aucune donnée d'entraînement n'est disponible. Seules les caractéristiques propres aux données de KBA sont ignorées dans le cadre de nos expériences sur ces données provenant de RepLab.

6. Se rapporter à la section 2.2.5

Le système a été utilisé avec les réglages suivants (i) les contenus textuels ont été nettoyés : suppression des mots-vides et symboles spéciaux tels que les « » avant les noms d'utilisateur et les «#» des mot-dièses, (ii) dans le cas où le message contient un lien, un ensemble de caractéristiques est également généré pour le document lié. Deux systèmes automatiques de classification ont été appris ; un premier à partir de l'ensemble des documents dits «*pertinents*» (alertes et importants) pour l'entité et «*non-pertinent*» (sans importance). Les documents semblant «*pertinents*» sont ensuite séparés par le deuxième système entre «*pertinents*» et «*très pertinents*» (alertes et importants).

Nous étudions l'impact du regroupement parfait (celui effectué manuellement par les annotateurs et validé par les experts superviseurs pour chaque entité) sur l'assignation du niveau de priorité (cf. tableau4.9) . Nous pouvons remarquer que rajouter l'information des groupes de messages de références permet d'améliorer de manière significative les systèmes de détection de priorité selon toutes les métriques. Cette observation prouve qu'une bonne définition des conversations [7] et groupes thématiques de messages permettrait d'affiner la détection des messages les plus pertinents.

En évaluant les comportements des méthodes au niveau de chaque classe, nous pouvons expliquer la contre-performance de la méthode «*KBA*» comme nous l'avons en section 4.1.3. Si cette dernière semble relativement robuste et précise lorsqu'il s'agit de détecter des contenus annotés comme «*non important*», elle semble également incapable de qualifier correctement les contenus les plus importants que les annotateurs ont considéré être des «*alertes*».

Nous proposons dans un second temps de combiner les sorties de nos systèmes de détection automatique de priorité avec les regroupements de messages issus de méthodes automatiques. En dehors du fait que l'ensemble des résultats montrés dans le tableau 4.10 sont inférieurs à ceux obtenus sans cette information de regroupements (cf. tableau 4.9) ; il est intéressant de noter qu'à l'exception de la fonction témoin, les performances des systèmes respectent la hiérarchie suivante : **F-m(baseline)** $<$ **F-m(MAP#1)** $<$ **F-m(Hierarch.)** $<$ **F-m(K-means)** $<$ **F-m(MAP#2)**. L'ordre dans lequel les combinaisons de systèmes sont classés respecte les performances en Reliability des méthodes de regroupement automatique (cf tableau 4.8).

7. Ou tout du moins la définition idéale proposée par les experts lors de l'annotation : des messages partageant un sujet commun ou très similaire.

Approche	F-Score			Précision	R.	S.	F(R,S)
	Alerte	Important	Sans Imp.				
Détection de priorité seule							
Lia_Prio_5	**.415**	**.684**	.646	**.627**	.387	**.315**	**.335**
KBA	.025	.560	**.705**	.585	.315	.276	.282
Baseline	.336	.643	.617	.530	**.403**	.248	.274
Détection de priorité avec les regroupements de référence							
Lia_Prio_5	**.514**	**.733**	.702	**.690**	**.549**	**.345**	**.387**
KBA	.002	.560	**.705**	.612	.532	.269	.329
Baseline	.441	.706	.703	.649	.511	.281	.326

TABLE 4.9 – *Performances des systèmes de détection de priorité sur les données d'évaluation de la collection RepLab 2013 ordonnées selon la métrique officielle du challenge : F-Measure (R,S) seuls puis en utilisant l'information des regroupements de référence.*

		F-Score						
	Approche	**Alerte**	**Important**	**Sans Imp.**	**Précision**	**R.**	**S.**	**F(R,S)**
Détection de priorité avec les regroupements de la baseline								
	Baseline	.336	.643	.617	.530	.403	**.248**	**.274**
	Lia_Prio_5	**.376**	**.672**	**.633**	**.550**	.520	.136	.172
	KBA	0	.478	**.661**	.489	**.578**	.071	.098
MAP Variante #1								
	Baseline	.342	.657	.659	.628	.383	**.151**	**.195**
	Lia_Prio_5	**.378**	**.660**	.646	**.632**	.413	.136	.181
	KBA	0	.466	**.672**	.568	**.551**	.098	.126
MAP Variante #2								
	Lia_Prio_5	**.373**	**.669**	.636	**.619**	.405	**.249**	**.288**
	Baseline	.329	.643	.628	.574	**.406**	.214	.261
	KBA	.069	.512	**.657**	.561	.361	.171	.217
Agglomération hierarchique basée sur l'indice de similarité Jaccard								
	Lia_Prio_5	.340	**.659**	.631	**.613**	.391	**.195**	**.239**
	Baseline	**.342**	.642	.631	.584	.378	.174	.214
	KBA	.126	.515	**.662**	.567	**.421**	.150	.192
Agglomération K-means basée sur l'indice de similarité Jaccard								
	Lia_Prio_5	**.365**	**.667**	.628	**.612**	**.416**	**.223**	**.269**
	Baseline	.338	.635	.625	.570	.392	.206	.253
	KBA	.130	.514	**.661**	.559	.409	.164	.212

TABLE *4.10* – *Performances des systèmes de détection priorité combinés avec les regroupements de messages obtenus avec des systèmes automatiques.*

Dans cette dernière expérience, nous étudions l'impact d'un regroupement imparfait obtenu automatiquement sur la référence manuelle de niveau priorité. Nous considérons la référence comme point d'entrée de notre protocole et nous essayons pour chaque groupe de messagess constitué par nos systèmes de propager le niveau de priorité majoritaire à l'ensemble de son groupe. Il est intéressant d'observer à partir des résultats résumés dans le tableau 4.11 dans quelle mesure les mauvais groupes dégradent la référence de priorité. Si les valeurs des F-Scores (précision/rappel) pour les niveaux Important et Sans-Importance restent comme pour la Reliability à des niveaux élevés, la chute de performance visible pour le niveau Alerte et encore plus manifeste sur la Sensitivity est inacceptable dans le cadre d'une application en production dans un cabinet d'analyse d'image de marque. Cette observation illustre d'une part la difficulté de la tâche à laquelle nous sommes confrontés et d'autre part l'immaturité (ou l'inadéquation au problème) des méthodes de classifications que nous avons proposées. En effet, ces dernières sont encore loin de pouvoir égaler le niveau d'interprétation d'un expert qui sait voir plus loin que la simple proximité lexicale entre deux messages.

Approche	F-measure			Précision	R.	S.	F(R&S)
	Alerte	**Important**	**Sans Imp.**				
Référence de priorité avec les regroupements de messages obtenus avec des systèmes automatiques							
MAP feat. select. #2	.710	**.840**	**.823**	**.812**	**.756**	**.518**	**.602**
Hierarch. clust.	.712	.785	.769	.783	.696	.438	.519
K-means clust.	**.769**	.815	.791	.761	.655	.367	.437
MAP feat. select. #1	.551	.754	.743	.731	.666	.229	.311
Baseline	.535	.763	.727	.634	.657	.198	.262

TABLE *4.11* – *Impact des regroupements de messages obtenus avec des systèmes automatiques sur l'annotation de priorité de référence.*

4.2.6 Conclusion

Nous avons étudié l'impact de la classification de messages sur la détection de niveaux de priorité. Nos expériences mettent en avant l'intérêt de cette hypothèse. Toutefois, les méthodes permettant aujourd'hui de regrouper les messages ne disposent pas encore de la maturité suffisante pour proposer des regroupements permettant améliorer les systèmes de catégorisation de priorité travaillant uniquement à l'échelle d'un message. Actuellement, cette succession d'approches amène à la propagation des erreurs commises dès la première étape.

4.3 Vers l'enrichissement des contenus

Nous avons vu dans ce chapitre que de nombreux participants à RepLab ont utilisé des ressources additionnelles pour enrichir le contenu des messages afin d'améliorer les performances de leurs systèmes. Nous proposons de comparer différents systèmes de contextualisation de tweets [8]. Cette comparaison est effectuée en utilisant l'informativité apportée par le contexte que fournissent ces méthodes pour affiner les systèmes de catégorisation thématique. Nous partons des résultats obtenus par un système de catégorisation thématique se basant uniquement sur le contenu des messages et nous cherchons à vérifier s'il possible de s'attendre à une amélioration des performances en considérant le message et son contexte plus ou moins réduit [9].

Il n'existe pas beaucoup d'études proposant d'évaluer l'informativité réelle de ces données contextuelles. Pendant plusieurs années, dans le cadre de CLEF le challenge « *INEX Tweet Contextualization* » [10] avait pour objectif d'automatiquement fournir une information contextuelle à partir de Wikipédia [11] afin de mieux comprendre un tweet. Les organisateurs de ce challenge souhaitaient également proposer un protocole permettant l'évaluation de ces contextes (Bellot et al., 2014).

Ces données représentent donc une opportunité incroyable pour tous ceux qui souhaitent s'intéresser à l'impact de l'ajout d'information sur les performances d'un système de catégorisation automatique. Nos expériences se focalisent autour des deux problématiques suivantes : (i) tout d'abord, est-ce que

8. Le contexte est obtenu à partir des liens hypertextes présents dans le message ou en interrogeant l'encyclopédie en ligne Wikipédia à partir des termes présents dans le message.
9. Entre une et cinq phrases
10. https://inex.mmci.uni-saarland.de/tracks/qa/
11. www.wikipedia.org/

cet enrichissement des données permet d'améliorer un système de catégorisation déjà très performant, (ii) existe-il un moyen d'évaluer automatiquement la qualité de ces informations contextuelles.

4.3.1 Enrichissement des contenus à partir des systèmes de contextualisation automatique

Nous utilisons l'information contextuelle produite par les soumissions officielles des systèmes automatiques de contextualisation ayant participé au challenge « *INEX Tweet Contextualization* 2014 ». Ces systèmes sont décrits en détails par leurs auteurs respectifs (Zingla M-A. et Y., 2014; Torres-Moreno, 2014; Ermakova, 2014).

Pour chaque message, chacun de ces systèmes propose un contexte se composant de différentes phrases plus ou moins pertinentes vis a vis du message. Nous proposons selon deux scénarios de sélectionner ces phrases pour enrichir le contenu des tweets que nous souhaitons catégoriser. Nous considérons les scénarios suivants :

— Sélection de la phrase la plus pertinente (selon le score fourni par le système de contextualisation) pour constituer un contexte bref de taille équivalente au message ;
— Sélection des cinq phrases les plus pertinentes pour constituer un contexte dit « *étendu* ».

Les documents ainsi enrichis sont soumis à un système de catégorisation thématique automatique avec l'espoir d'améliorer la catégorisation. Nous proposons également de soumettre le contexte long sans le message d'origine à ce même système de catégorisation afin d'estimer s'il est possible de remplacer un message par son contexte.

Nous considérons deux variantes du système automatique cosine que nous avions proposé lors de la tâche de catégorisation thématique RepLab 2014, une version considérée comme une « *baseline faible* » et une autre qualifiée de « *forte* » ou « *optimisée* » pour cette tâche [12]. Ces variantes sont notées « *weak* » et « *strong* » baselines sur les figures 4.3, 4.4 et 4.5. Sans enrichissement, sur les 77 messages de la sélection, la variante dite « *faible* » s'illustre par des performances de catégorisation logiquement plus faibles en obtenant un « *F-Score* » de .56 et une « *Accuracy* » à .60 (contre .65 et .71 pour la version optimisée) [13].

12. L'optimisation a été conduite à partir de l'ensemble de développement que nous avions constitué avec l'ensemble d'entraînement.
13. Nous verrons d'ailleurs en section 4.3.4 que la contextualisation ne permet pas d'améliorer ces scores.

D'ailleurs ces performances plus élevées que celles obtenues par la même méthode sur l'ensemble de la collection d'évaluation liée à cette tâche de RepLab 2014 indiquent sans doute que ces contenus sélectionnés pour leur lisibilité et leur clarté sont traités plus facilement par un système automatique.

4.3.2 Enrichissement des contenus à partir du système de généralisation lexicale

Nous avons entraîné un modèle Word2Vec(Mikolov et al., 2013a) de skip-gram (jusqu'à 10 mots de contexte) de 600 dimensions pour généraliser à la fois des termes en anglais et en espagnol. Ce modèle a été entraîné à partir des messages non-annotés fournis par les organisateurs de RepLab (Amigó et al., 2013a) auxquels nous avons rajouté des corpus facilement accessibles [14].

Ce modèle est ensuite utilisé comme modèle de généralisation. Il a pour objectif de remplacer chaque mot hors vocabulaire par une sélection des N mots les plus proches dans cette représentation distribuée continue des mots selon deux contraintes (i) sa présence dans le vocabulaire connu dans l'ensemble d'entraînement (ii) une pureté G_i (2.3) importante.

4.3.3 Comparaison de systèmes de contextualisation automatique

La figure 4.2 montre les résultats officiels du challenge « *INEX Tweet Contextualization* 2014 » selon les critères d'informativité et lisibilité chaque système est identifié par un entier (*3xx*). Les organisateurs du challenge (Bellot et al., 2014) considèrent qu'il est difficile de sélectionner un passage de wikipédia donnant à la fois une information contextuelle sur le message et à la fois de l'information liée à l'image de marque de l'entité.

Il apparaît également sur la figure 4.2 qu'il existe dans l'état de l'art un compromis entre lisibilité et informativité. Les systèmes privilégiant l'une de deux mesures étant très mal classés selon l'autre. Le meilleur système selon le critère d'informativité est basé sur une association de règles de décision (Zingla M-A. et Y., 2014). Selon l'autre critère, c'est un résumeur automatique avancé (Torres-Moreno, 2014) veillant à inclure un minimum de modifications aux passages extraits de Wikipédia qui obtient les meilleurs performances en terme de lisibilité.

14. enwik9, One Billion Word Language Modelling Benchmark, the Brown corpus, English GigaWord 1 à 5, eswik, parallel es-en europarl.

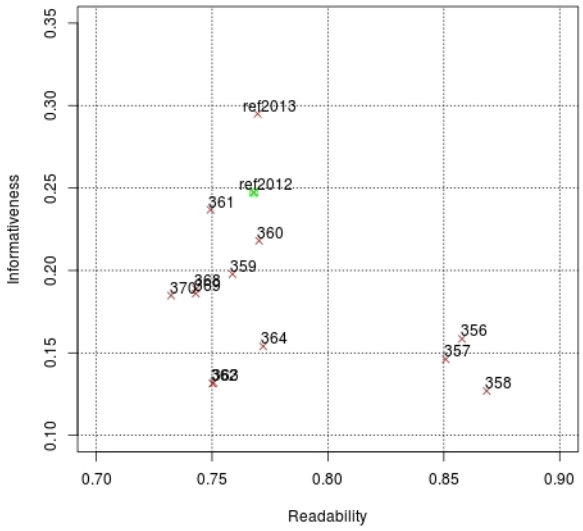


FIGURE 4.2 – *Comparaisons des performances des participants au challenge INEX Tweet Contextualization* 2014 *selon la lisibilité et l'informativité.*

4.3.4 Évaluation : cas de la catégorisation thématique

Nous utilisons les données et les mesures d'évaluation relatives à la tâche de catégorisation thématique du challenge RepLab 2014 décrites dans le chapitre 3. A noter, sur les 240 messages sélectionnés dans le cadre du challenge « *INEX Tweet Contextualization* 2014 » seulement 77 messages sont présents dans la collection d'évaluation RepLab 2014. Par rapport à la distribution initiale des thématiques dans les ensembles d'entraînement et évaluation vus dans le tableau 3.4 notons à titre indicatif que la thématique majoritaires « *produits et services* » ne représente que 30% des messages contre presque 50%.

Utilisation de la contextualisation

La figure 4.3 montre les performances de catégorisation thématique en substituant le contexte « *étendu* » des messages au contenu textuel de ces derniers. Le premier groupe d'indicateurs de performances (les quatres barre les plus à gauche) rappelle le niveau de performance du système de catégorisation thématique sur ces mêmes messages sans contexte. Nous pouvons noté qu'aucun contexte ne permet d'égaler le niveau de performance obtenu à partir du contenu textuel originel des messages.

La figure 4.4 présente les performances de catégorisation thématique à partir

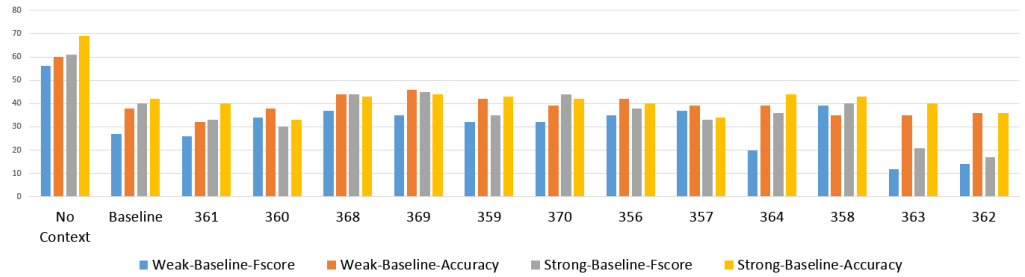


FIGURE 4.3 – *Performances de catégorisation thématique sur les 77 messages de la collection en se basant uniquement sur le contexte dit « étendu » à cinq phrases.*

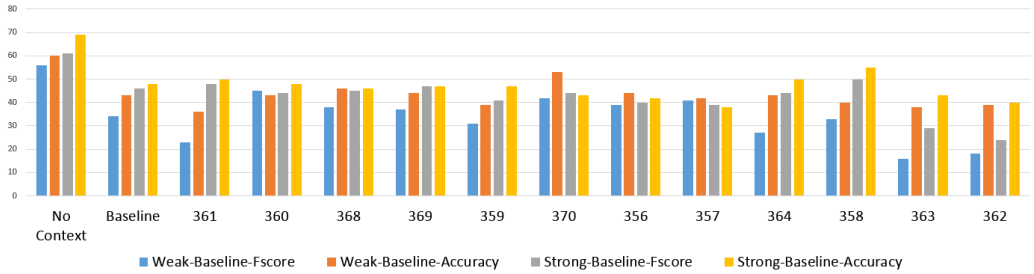


FIGURE 4.4 – *Performances de catégorisation thématique en considérant le contexte « étendu » en complément du contenu textuel.*

du contenu textuel des messages complété par son contexte étendu. Les résultats des différences combinaisons sont légèrement plus élevé que ceux observés précédemment cependant la performance « *baseline* » obtenue sans contexte additionnel reste inégalée.

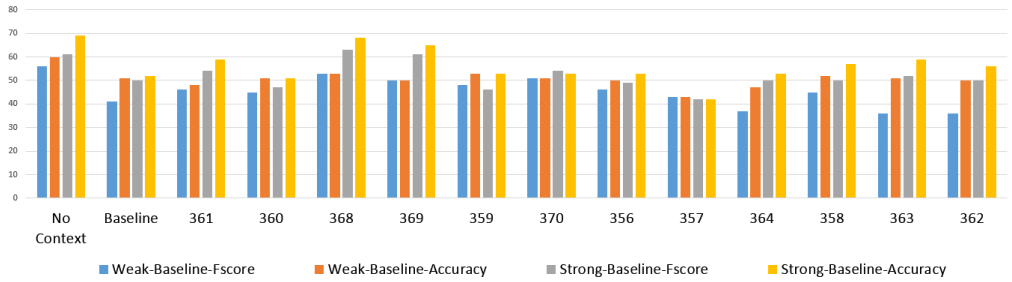


FIGURE 4.5 – *Performances de catégorisation thématique en considérant le contexte court en complément du contenu textuel.*

Après avoir observé l'inefficacité du contexte étendu que peut-on attendre de son équivalent court ? La figure 4.5 montre les performances de catégorisation thématiques sur les 77 messages de la collection en considérant les différents contextes courts, proposés par les systèmes automatiques de contextualisation ayant participés à INEX 2014, en complément du contenu textuel des

messages. Cette fois-ci les systèmes de contextualisation automatiques INEX 368 et 369 montrent des améliorations des systèmes automatiques de catégorisation thématique. Cette amélioration indique que le contenu additionnel du contexte apporte de l'information pertinente pour la catégorisation sans rajouter de bruit comme ce fut le cas des contextes étendus.

Soumissions INEX	Lisibilité	Informativité	Classification
358	1	11	6
356	2	8	10
357	3	9	13
364	4	10	10
360	5	3	12
Baseline	6	1	9
359	7	6	11
363	8	12	4
362	9	13	7
361	10	2	3
368	11	4	1
369	12	5	2
370	13	7	5

TABLE 4.12 – *Rang des soumissions officielles au challenge INEX Tweet Contextualization* 2014 *selon l'informativité, la lisibilité et l'apport d'information dans la tâche de catégorisation thématique.*

A partir de cette dernière expérience, nous proposons dans le tableau 4.12 un nouveau classement des soumissions au challenge INEX « *INEX Tweet Contextualization* 2014 » à partir du gain obtenu par le contexte qu'ils proposent dans la tâche de catégorisation thématique. Les deux classements officiels basés sur l'informativité et la lisibilité sont totalement bouleversés. A l'instar du classement en informativité, cette fois encore les systèmes ayant obtenu de bons résultats de lisibilité semblent une fois encore pénalisés.

Notre objectif était d'étudier si l'enrichissement à partir du seul texte du wikipedia peut améliorer une tâche objective de catégorisation thématique de tweets. Si les 77 tweets d'INEX étaient suffisant pour mener les premières expérimentations, ils demeurent largement insuffisant pour tirer des conclusions généralisables et robustes.

Nous pouvons conclure cette série d'expérimentations sur le fait que les méthodes automatiques que nous avons proposées et plus généralement celles décrites dans la littérature sont bridées par la qualité et la quantité de données auxquelles elles sont confrontées. Comme nous l'avons observé, à l'instar des experts humains, ces méthodes sont beaucoup plus à l'aise avec des messages

pour lesquels le contenu textuel semble de meilleure qualité (absence d'URL et nombre de mots plus important).

Utilisation de la généralisation lexicale : évaluation

Les performances des systèmes de catégorisation thématique sont comparées sur la collection d'évaluation selon le « *F-Score* » et « *l'accuracy* » dans le tableau 4.13. Nous avons déjà constaté en section 4.1.1 que les fonctions témoins proposées par les organisateurs de RepLab 2014 (Amigó et al., 2014) étaient complètement dépassées par l'ensemble des systèmes. Nous avons également signalé que les systèmes (McDonald et al., 2015) et (Rahimi et al., 2014) utilisaient des méthodes d'enrichissement des contenus, basées sur « *les documents pseudo-pertinents* » pour les premiers et sur une méthode « *expansion sémantique* » pour les seconds. Nous proposons de revenir sur notre double évaluation de la catégorie « *Non Définie* ». Cette fois-ci en enrichissant les contenus des messages de la collection d'évaluation (systèmes notés « *w/ Context* » dans le tableau 4.13).

Approche	F-Score (-U)	Précision (-U)	F-Score (+U)	Précision (+U)
cosine w/ Context	**.505**	.739	.494	.707
cosine	.491	.736	.500	.693
SVM	.469	.732	.461	.679
SVM w/ Context	.468	.732	.456	.679
(McDonald et al., 2015)	.473	.731	-	-
(Rahimi et al., 2014)	.489	.695	-	-
SVM Baseline	.38	.622	-	-
Naive Baseline	.152	.560	-	-

TABLE 4.13 – *Performances des systèmes automatique de catégorisation thématiques sur les données Replab* 2014 *ordonnées par Accuracy (colonne Acc(-U)). Les méthodes utilisant la généralisation lexicale sont notées (w/ Context). Les scores les plus élevés sont indiqués en gras, les améliorations significatives (sur la moyenne par entité) par rapport au système SVM -U (t-test appairé $p < 0:05$) marqués par une *.*

Le système cosine est significativement [15] amélioré par l'ajout du contexte lexical. Les CRF sont également renforcés par cette information supplémentaire alors qu'elle n'a aucun effet sur la performance des SVM. D'ailleurs ces méthodes basées sur des SVMs (normalement les plus discriminantes) sont les plus perturbées par la nouveauté et la variabilité apportées par la catégorie « *Non Définie* ».

4.3.5 Évaluation : cas de la généralisation thématique appliquée à la détection de priorité

A l'instar de l'expérience précédente, nous considérons les mêmes systèmes de catégorisation automatique appliqués cette fois à la question de la détection automatique de priorité. Nous comparons dans le tableau 4.14 les performances des systèmes avec et sans généralisation lexicale. Quelle que soit la métrique (F-Score moyen par classe, Précision ou F(R,S) moyenne sur l'ensemble des entités), les différences entre ces variantes d'une même méthode ne sont pas significatives. Cette observation indique que contrairement à la question de la catégorisation thématique, le contexte lexical ne permet pas d'ajouter une information porteuse des caractéristiques implicites de priorité d'un message.

Approche	F-Score	Précision	F(R,S)
(Cossu et al., 2013) *Lia_Prio_5*	**.571**	.636	**.335**
SVM w/ Context	.564	.645	.304
SVM	.563	.644	.304
(Cossu et al., 2013) *KBA*	.421	.585	.282
Baseline	.512	.570	.274
cosine w/ Context	.562	.634	.260
cosine	.561	.633	.260
(Cossu et al., 2013) *Fusion PROMETHEE*	-	.651	.253
(Cossu et al., 2013) *Combinaison Linéaire*	-	.647	.251
(Cossu et al., 2013) *Fusion ELECTRE*	-	**.652**	.251
Median	-	*.573*	*.249*

TABLE 4.14 – *Performances des systèmes de détection de priorité sur les données d'évaluation de la collection RepLab* 2013 *ordonnées selon la métrique officielle du challenge : F(R,S). Les méthodes utilisant la généralisation lexicale sont notées « w/ Context ».*

15. Test de significativité standard : « *t-test appairé* » avec $p < 0:05$

4.4 Conclusion et perspectives

Nous avons déjà constaté que les performances des approches statistiques de TAL sont limitées par la quantité et la qualité des données textuelles à leur disposition. Les expériences menées dans le cadre de ce chapitre nous permettent de dire qu'à partir du moment où il y a suffisamment de données d'apprentissage, le contenu lexical des messages est suffisant pour apprendre efficacement à distinguer automatiquement les différentes catégories dont relèvent les messages. Par ailleurs, comme nous pouvons le retrouver dans la littérature, nos expériences mettent en avant la pertinence de la contextualisation (lexicale ou thématique) pour les systèmes automatiques nécessitant plus de données ou pour compenser le manque d'information pour les catégories ne comportant que peu d'exemples. Cet enrichissement des contenus textuels peut facilement s'effectuer à partir de grandes collections de Micro-Blogs de même nature ou bien en utilisant des collections de documents de meilleure qualité comme Wikipiédia ou le ClueWeb B. Sur des contenus plus longs et moins susceptibles d'être vagues, l'intérêt de la contextualisation est discutable selon l'application. Si la contextualisation ajoute de l'information désambiguïsant le message et permettant d'affiner la détection d'une thématique, le contenu additionnel est généralement «*neutre*» en termes d'opinion (polarité) et priorité et son utilisation peut se révéler contre productive.Enfin, notre analyse des sorties de contextualisation dans le cadre du challenge «*Inex Tweet Contextualisation*» révèle un lien entre l'informativité des phrases et leur utilité pour étendre les messages concernant l'image de marque d'une entité. Cette utilité mesurée au travers des gains de performances dans la tâche de catégorisation thématique représente une évaluation indirecte de l'informativité.

Chapitre 5

Profilage d'utilisateur

Sommaire

5.1 Introduction

5.1.1 Le profilage, mais pour quoi faire ?

Nous nous sommes intéressés dans le chapitre 4 à la détection des opinions et thématiques véhiculées dans les messages publiés par les utilisateurs de différents médias sociaux. Mais, dans le fond, qui sont vraiment ces utilisateurs ? Notre besoin de suivre de l'image de marque d'entités sur le Web 2.0 implique également de mieux connaître les individus qui mentionnent les entités mais aussi et surtout ceux qui contribuent à véhiculer l'image de marque dans le réseau, ces derniers sont souvent qualifiés de : « *leaders d'opinions* » ainsi ne s'intéresser qu'aux messages les plus importants.

Cette analyse est appelée le « *profilage* », elle a pour objectif de mieux comprendre le comportement des utilisateurs et d'identifier des groupes dont certains sont plus ou moins importants que les autres. En raison de sa popularité et de sa large utilisation, nous nous arrêtons cette fois-ci plus particulièrement sur le réseau social Twitter. Beaucoup de raisons poussent toutes sortes d'entités à catégoriser les utilisateurs de ce réseau, citons par exemple les entreprises qui visent une population type pour redorer leur image, les journalistes à la recherche des meilleures sources d'informations, les services de renseignement qui cherchent à identifier des personnes (ou groupes) potentiellement dangereuses, les services de police et médicaux qui recherchent les personnes suicidaires (Abboute et al., 2014) mais aussi, les services météorologiques qui cherchent à détecter et relayer de l'information pour prévenir certaines populations de phénomènes naturels dangereux (Sakaki et al., 2010) ou encore les épidémiologistes qui cherchent à prévoir l'expansion d'une maladie (Sadilek et al., 2012). Citons enfin la politique, où beaucoup d'entités s'intéressent aux opinions des utilisateurs pour prédire le résultat des élections (Park et al., 2011) et (Sobkowicz et Sobkowicz, 2012). Toutefois, la problématique du profilage des utilisateurs peut être abordée selon différents angles, en termes d'orientation, de capacité à être influencé ou peut encore viser à catégoriser l'utilisateur selon des catégories socioprofessionnelles (CSP, âge, humeur, etc.). C'est d'ailleurs pour y voir plus clair dans cette variété d'approches que beaucoup d'études se sont concentrées sur la caractérisation du profil des utilisateurs.

Twitter étant avant tout un service de diffusion d'informations, un grand nombre de chercheurs a essayé d'identifier les utilisateurs ayant un rôle particulier au coeur du service. La tâche de détection de « *spammeurs* » est l'une des plus populaires à sujet. Nous pouvons mentionner de récents travaux s'intéressant à l'identification « *robots-spammeurs* » (Benevenuto et al., 2010; Lee et al., 2010; Wang, 2010; Lee et al., 2011; Ghosh et al., 2012) et leurs équivalents humains

« *crowdturfers* » [1] (Chu et al., 2012; Lee et al., 2013). Nous pouvons également citer les travaux visant à repérer les « *capitalistes sociaux* ». Il s'agit d'un type particuliers d'utilisateurs cherchant au travers de plusieurs stratégies à gagner de la visibilité sans pour autant produire des contenus informatifs ou ayant une quelconque utilité (Ghosh et al., 2012; Dugué et Perez, 2014; Dugué et al., 2015). D'autres travaux sont allés plus loin dans la catégorisation de ces nouveaux « *acteurs du numérique* » en proposant de distinguer les « *vrais utilisateurs* » (individus, professionnels, entités) des « *services* » (spammeurs, flux de nouvelles, services marketing et commerciaux) (Uddin et al., 2014).

La problématique de la détection d'influence est également un sujet d'actualité (Anger et Kittl, 2011; Weng et al., 2010; Cossu et al., 2015b), avec une tâche de conférence (CLEF) et des ateliers (comme SocInf [2]) qui lui sont totalement dédiés. L'objectif de ces travaux est principalement d'établir une métrique permettant de mesurer l'influence des utilisateurs ou de chercher directement à détecter ceux qui sont influents parmi la masse de contributeurs. D'autres équipes proposent de catégoriser les utilisateurs du réseau social selon des aspects plus conventionnels comme les catégories socioprofessionnelles ou le domaine d'activité, l'âge (Al Zamal et al., 2012; Rangel et al., 2014; Rao et al., 2010) et le genre (Al Zamal et al., 2012; Rangel et al., 2014; Rao et al., 2010), l'origine ethnique ou géographique (Pennacchiotti et Popescu, 2011; Rao et al., 2010; Cheng et Lee, 2010; Mahmud et al., 2012; Huang et al., 2014), l'orientation politique (Al Zamal et al., 2012; Conover et al., 2011; Makazhanov et Rafiei, 2013; Pennacchiotti et Popescu, 2011; Rao et al., 2010) ou encore le type d'utilisateur (les organisations d'un côté et les individus qui peuvent être séparés en journalistes et personnes ordinaires de l'autre) (Pennacchiotti et Popescu, 2011; de Silva et Riloff, 2014; de Choudhury et al., 2012).

Comme ces études proviennent toutes de domaines parfois très éloignés comme : informatique, sociologie, statistiques, sciences politiques, elles ont forcément des objectifs très différents et investiguent chacune la question à leur manière en manipulant leurs propres données et en y appliquant leurs méthodes. Elles possèdent toutefois un point commun sur la manière d'aborder le sujet. La première étape à partir de données collectées sur le média social reste toujours la recherche et l'identification d'une caractéristique particulière permettant de décrire un utilisateur de façon à répondre correctement au problème abordé. Beaucoup de ces caractéristiques sont, d'ailleurs, spécifiques aux domaines de recherche. Les travaux en analyse des réseaux sociaux s'intéressent aux interconnexions entre utilisateurs alors que les travaux en traitement automatique de la langue se focalisent sur les contenus publiés par les utilisateurs en question. Il existe bien sûr des caractéristiques (le plus souvent très simples)

1. Personnes payées pour avoir des comportements de « *robots-spammeurs* »
2. http://socinf2015.isistan.unicen.edu.ar/

universellement utilisées comme le nombre de messages publiés, le nombre de relations de l'utilisateur (abonnés, abonnements, amis). C'est ensuite à partir de ces caractéristiques que sont appris des systèmes de classification automatique pour séparer les utilisateurs intéressants des autres. A ce titre, nous proposons une liste des caractéristiques qu'il est possible d'extraire de Twitter afin de caractériser un utilisateur.

Du fait de leur variété, il reste toutefois difficile de pouvoir déterminer avec certitude l'intérêt d'une caractéristique pour répondre à une problématique donnée, ou de leur pertinence lorsqu'il s'agit d'appréhender des questions plus spécifiques. La littérature est également floue au sujet de leur dénomination chacun le faisant avec les «*a priori*» de son domaine d'étude et d'application. Une même caractéristique peut être retrouvée dans plusieurs travaux sous différentes dénominations de même, il est possible de retrouver des caractéristiques différentes sous une même dénomination. Nous faisons l'impasse sur toutes les manières différentes sous lesquelles une caractéristique peut être définie (que l'on peut exprimer par tweet ou par profil, avec valeurs min, max, moyenne et écart-type) comme l'ont souligné (Cossu et al., 2015b). Les organisateurs des dernières éditions de PAN[3] n'ont d'ailleurs toujours pas réussi à faire ressortir une caractéristique qui répondrait à la problématique de détection d'âge et genre (Rangel et al., 2015).

5.1.2 Notion d'influence

Le *dictionnaire Larousse* définit l'influence comme «*pouvoir social et politique de quelqu'un, d'un groupe, qui leur permet d'agir sur le cours des événements, des décisions prises, etc.*». S'agissant de l'influence d'un utilisateur donné sur Twitter, beaucoup de facteurs entrent alors en ligne de compte.

Il existe de nombreuses méthodes pour l'analyse de l'orientation et l'influence d'un utilisateur. Beaucoup considèrent intuitivement que l'influence est liée à l'audience. Plus l'utilisateur est suivi, mentionné et repris, plus il est «*influent*». Un second indicateur, tout aussi utilisé, repose sur l'exploitation la structure du réseau social interconnectant les utilisateurs, partant du postulat qu'un lien, une mention est assimilable à un vote de l'utilisateur u_x en faveur de l'autorité de l'utilisateur u_y. Certaines recherches parlent d'une mesure d'autorité relative pour tous les membres du réseau. Des algorithmes comme le PageRank permettent de calculer un score pour chaque utilisateur, ce score modélise la probabilité qu'une personne navigant aléatoirement à travers la structure

3. Challenge de détection de plagiat, identifiant et profilage d'utilisateurs de médias sociaux. Plus d'informations sur http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/

du réseau social visite le profil de cet utilisateur. Les utilisateurs vers lesquels le plus de monde converge sont considérés comme ceux ayant la plus grande autorité. Ces mesures introduisent également un biais en considérant que le fait de s'abonner à un utilisateur influent augmente notre influence.

En tirant parti de ces lacunes, certains utilisateurs appelés «*capitalistes sociaux*» ont d'ailleurs fait l'objet de récentes études. Le «*capitalisme social*» est une stratégie visant à maximiser de manière artificielle son influence en se connectant à un grand nombre d'utilisateurs pratiquant eux-mêmes cette stratégie dans l'espoir que s'établisse un lien réciproque. Ces utilisateurs gagnent ainsi en visibilité sur le réseau sans avoir produit un quelconque contenu de qualité. Cette pratique remet ainsi en cause les méthodes que nous venons juste d'évoquer. Mais comme nous avons pu le voir précédemment, (Dugué et Perez, 2014) ont constaté une caractéristique propre à ces utilisateurs : le taux de recouvrement important dans leurs interactions sur le réseau (voisinage entrant et sortant).

Durant l'atelier SocInf, beaucoup de chercheurs ont essayé de combiner plusieurs caractéristiques sous la forme de catégories comme les interactions, le profil public et les contenus produits. Toutefois, il n'y a pour l'instant aucun consensus au sujet des caractéristiques les plus pertinentes qu'il convient d'employer à des fins analytiques, ni même celles qui seraient au minimum efficaces pour caractériser l'influence d'un utilisateur. Il existe des études comme celle de (Kim et al., 2014) indiquant directement comment interagir et publier du contenu afin de maximiser sa visibilité et son influence [4]. De notre point de vue ces études amènent certains utilisateurs à suivre des comportements pouvant induire en erreur ou «*fausser*» les méthodes d'analyse d'influence.

Dernière question et pas des moindres : comment l'influence mesurée sur Twitter (ou sur n'importe quel autre média social) se traduit-elle en terme de véritable influence ? Quelques chercheurs ont proposé des méthodes permettant de détecter les «*influents*» (parfois également appelés «*leaders d'opinions*») sur le réseau. Toutefois en dehors de quelques rares personnalités reconnues comme étant influentes, la validation reste impossible. Deux études ont exploré la question pour le réseau social Facebook en étudiant comment nos «*amis*» influencent notre comportement durant les élections. Ces études montrent que l'abstention est liée à la connaissance de l'action de vote de nos amis. En clair nous sommes plus incités à aller voter si tous nos amis nous indiquent qu'ils sont allés voter. Cette influence que l'on peut qualifier «*d'influence en ligne*» est appelée «*influence sociale*» dans certains travaux comme ceux de (Anagnostopoulos et al., 2008) et se distingue de «*l'influence réelle*» (également dite «*hors ligne*») qui correspond à l'influence de la personne constatée hors d'un

4. http://goo.gl/XFnBD7

média social.

5.2 Définition d'un profil

Tous les médias sociaux sont des services en ligne proposant à l'utilisateur, de créer une page de profil à partir de laquelle ils peuvent avoir une activité de publication, se connecter à d'autres utilisateurs pour suivre leurs publications et interagir sur ces dernières. Chaque service se distinguant par la suite avec ses spécificités et son vocabulaire mais également au travers de sa visibilité et de l'accessibilité des pages de profil de leurs utilisateurs (et donc des informations plus ou moins précises que l'on trouve au sujet de ces derniers). Par exemple, contrairement à Facebook, les profils d'utilisateurs Twitter sont, par défaut, accessibles à tout un chacun sans avoir à posséder un compte Twitter et indexés par les moteurs de recherche traditionnels. Autre spécificité de Twitter, les liens entre utilisateurs sont unilatéraux ce qui sous-entend que n'importe quel utilisateur peut se connecter à n'importe quel autre.

Bien sûr, comme il est impossible d'être exhaustif étant donné le nombre de travaux existants en matière de caractérisation de profil d'utilisateur. Nous nous limiterons dans cette partie à ne parler que des caractéristiques les plus intéressantes. Comme il n'existe pas de standard définissant et regroupant les différentes caractéristiques permettant de définir un profil nous nous baserons sur la catégorisation que nous avons déjà proposée (Cossu et al., 2015b). Nous posons à partir de l'état de l'art, qu'un profil Twitter peut être défini à partir des éléments suivants :

— un profil public ;
— une activité de publication ;
— un réseau de relations ;
— des interactions avec ce réseau ;
— un champ lexical ;
— un style éditorial ;
— toutes sortes de données externes.

Nous décrivons plus en détail dans les sections suivantes chacune des catégories que nous venons d'évoquer.

5.2.1 Profil Public

Le profil public est la page vitrine de l'utilisateur, c'est souvent la première contribution que l'on voit de l'utilisateur et parfois la dernière si celle-ci ne contient aucune information pertinente. La figure 5.1 montre la page de profil

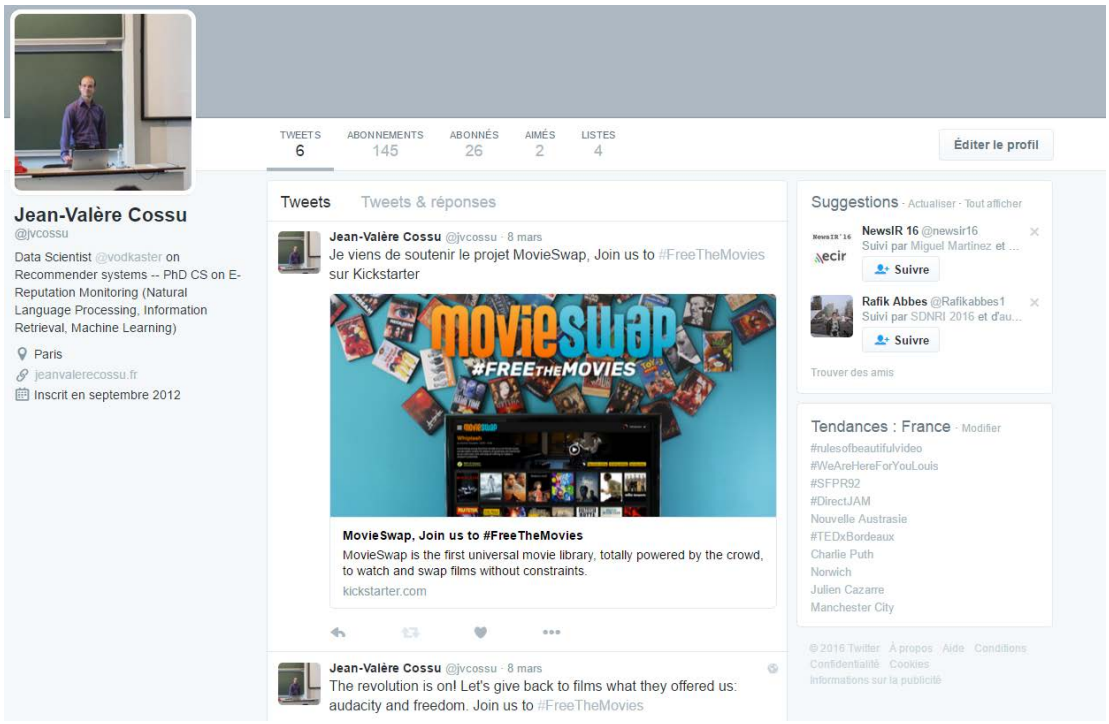


FIGURE 5.1 – *Exemple de profil public d'un utilisateur de Twitter.*

d'un utilisateur de Twitter. Nous définissons quatre indicateurs booléens indiquant, la présence d'un avatar (image de profil), la certification du compte par Twitter (contrôlé par son propriétaire légitime), l'autorisation des contributions (dans le cas où un même compte est partagé par toute une équipe éditoriale) et enfin la présence d'un lien vers une page web personnelle. Nous considérons également l'auto-description fournie par l'utilisateur, sa longueur, les éléments potentiellement présents comme les liens et mentions d'autres utilisateurs (certains utilisateurs précisant qu'ils ont plusieurs comptes ou sites). Notons que certains chercheurs (Huang et al., 2014) ont été jusqu'à analyser l'avatar avec des méthodes d'analyse d'images pour récupérer plus de détails sur l'utilisateur comme l'âge, le genre et l'origine ethnique. A ce titre, la présence de caractères spéciaux et le nom sont également considérés comme de bons indicateurs (Pennacchiotti et Popescu, 2011; Huang et al., 2014).

Ces caractéristiques donnent finalement de bonnes indications sur l'attention que porte l'utilisateur à la perception que les autres utilisateurs peuvent se faire de sa présence en ligne. Généralement, les comptes de personnalités et grandes entreprises sont certifiés (Chu et al., 2012) et l'ensemble des champs sont remplis correctement là où les « *robots* » les laissent vides. D'ailleurs, la date de création du compte est également une donnée importante, (Chu et al., 2012) rapportent en effet que la plupart des « *robots* » ont été inscrits en 2009 sur le

réseau social. Le partenariat signé avec Google à cette date à sans doute amener un meilleur algorithme de détection de ce type d'utilisateurs lors de leur inscription.

5.2.2 Activité de publication

S'il y a un bien un point où les médias sociaux diffèrent particulièrement des médias traditionnels c'est l'activité de publication. Avec la limite des 140 caractères, les messages courts publiés sur Twitter sont expressifs et très denses. En outre, avec la guerre des utilisateurs à laquelle se livre les différents réseaux sociaux, Twitter compte un nombre important d'utilisateurs actifs publiant jusqu'à 400 millions de messages par jour alors que les médias traditionnels reposent sur un nombre restreint de contributeurs et de contributions. Enfin, hormis certains cas particuliers il n'y a pas de filtre sur les contenus publiés, contrairement aux médias traditionnels, les médias sociaux ont introduit le régime de l'immédiateté de la publication ce qui amène parfois à quelques ratés de communication comme cela a pu arriver, il y a encore peu de temps, lorsqu'une journaliste de la BBC annonçait par erreur le décès de la reine d'Angleterre [5].

Tout particulièrement, nous nous intéressons ici à la manière dont les utilisateurs publient leurs messages. La première variable qu'il semble important de regarder est naturellement le nombre de messages publiés, cette variable ayant une forte amplitude de valeur. Afin d'évaluer la régularité de l'utilisateur beaucoup de chercheurs comme (Java et al., 2007; Benevenuto et al., 2010; Lee et al., 2011; Pennacchiotti et Popescu, 2011) se sont mis à considérer des variantes avec les publications par jour, mois etc. Cette évaluation de la périodicité des publications permet également de détecter les «*bots*» [6] programmés pour tweeter à intervalles réguliers (Chu et al., 2012). La possibilité de joindre d'autres médias (photos, vidéos) aux messages a également fait l'objet de nombreuses études. Les utilisateurs travaillant dans le domaine de l'image (photographie, télévision, cinéma) associent généralement plus souvent des photos à leurs messages.

Avec les avancées des technologies, Twitter offre la possibilité de géolocaliser les utilisateurs et plus particulièrement de géolocaliser individuellement chaque message. Certains auteurs comme (Vilares et al., 2014) considèrent que le simple fait d'activer cette option est discriminant alors que d'autres comme (Huang et al., 2014) utilisent la localisation associée au message pour déterminer des thématiques associées aux messages. Cette fonctionnalité est très appréciée des personnes voulant avoir un contrôle total de leur image même si

5. http://goo.gl/xJCrgG
6. Agent automatique ou semi-automatique qui imite le comportement d'un humain.

ce n'est parfois qu'une illusion. Dans certains cas, cette fonctionnalité est utile pour différencier les messages postés depuis son lieu de travail ou de vacances.

5.2.3 Réseau de relations

Nous nous intéressons ici à la manière dont l'utilisateur est relié au réseau Twitter. Ce réseau est pour rappel principalement basé sur la relation unilatérale abonnement-abonné. Ces deux valeurs sont d'ailleurs considérées séparément car elles illustrent deux comportements distincts. Le nombre d'abonnés indique à quel point ce que peut publier l'utilisateur intéresse les autres utilisateurs. Le nombre d'abonnements de l'utilisateur indique à l'inverse à quel point l'utilisateur s'intéresse aux publications des autres utilisateurs. Pour les chercheurs en analyse des réseaux sociaux, ces valeurs correspondent aux degrés entrants et sortants du nœud représentant l'utilisateur dans le réseau. L'autre mesure la plus utilisée à partir de ces deux variables est le ratio abonnés-à-abonnements (Benevenuto et al., 2010; Rao et al., 2010; Wang, 2010; Al Zamal et al., 2012; Lee et al., 2013).

Ces relations au réseau se révèlent être instructives. C'est à partir d'elles que (Dugué et Perez, 2014) ont réussi à distinguer les fameux «*capitalistes sociaux*» du reste des utilisateurs. Ces utilisateurs ayant la particularité de tirer parti de chaque spécificité du réseau pour augmenter leur visibilité sur Twitter. Une conséquence de leur stratégie est le fort recouvrement entre leurs abonnements et abonnés, créant ainsi une bulle à part d'utilisateurs fortement liés entre eux. Dans leurs travaux, (Dugué et Perez, 2014) indiquent qu'il suffit de s'intéresser aux 5000 abonnés et abonnements les plus récents pour calculer ce taux de recouvrement. Les «*robots-spammeurs*» utilisent également cette stratégie d'auto abonnements croisés les uns aux autres pour accroître leur visibilité. Ces comptes étant créés rapidement à la chaîne, l'aspect consécutif de leurs identifiants est également un marqueur permettant de les repérer. De cette façon, il est aussi possible d'épingler quelques personnalités souhaitant visiblement se montrer plus populaires et appréciées qu'elles ne le sont vraiment. Enfin, s'intéresser au nombre de messages publiés par les abonnés ou abonnement d'un compte permet également de vérifier le niveau d'activité de ces derniers et ainsi vérifier (pour les abonnements) la légitimité de l'intérêt que l'on peut leur porter.

Plutôt que de se concentrer sur cette topologie locale, d'autres chercheurs s'intéressent au réseau dans sa globalité. Par exemple, (Weng et al., 2010) ont proposé une modification de l'algorithme de «*PageRank*» pour calculer une influence restreinte à un sujet de conversation. (Java et al., 2007) ont utilisé les mesures de centralité (hub vs. authorities) pour détecter des communautés d'uti-

lisateurs autour d'un sujet et les utilisateurs les plus intéressants au sein de ces communautés.

Via les listes, Twitter propose de regrouper les abonnements selon ses convenances puis de partager ces listes avec d'autres utilisateurs, le nombre de listes dans lesquelles un utilisateur donné est présent est un bon indicateur de l'intérêt porté à ses publications.

5.2.4 Interactions avec le réseau de relations

Les interactions rassemblent les caractéristiques qui définissent la façon dont un utilisateur interagit avec les autres utilisateurs. Le principal élément d'interaction avec les autres utilisateurs est la possibilité de reprendre la publication d'un autre utilisateur et de la (re)diffuser auprès de son propre réseau, généralement pour montrer son accord (Boyd et al., 2010). La proportion de reprises parmi les messages d'un utilisateur, c'est à dire le nombre de messages qu'il publie alors qu'il n'en est pas l'auteur, est une des variables les plus utilisées à ce titre (Benevenuto et al., 2010; Rao et al., 2010; de Choudhury et al., 2012).

Ici encore, la variable est considérée selon différentes périodes temporelles (Uddin et al., 2014). A l'inverse, nous pouvons également considérer la variable complémentaire, c'est à dire la proportion de tweets de l'utilisateur qui sont repris par les autres utilisateurs, indiquant alors à quel point les publications de l'utilisateur intéressent le reste du réseau. Certaines études (Anger et Kittl, 2011) ne s'intéressent qu'aux publications ayant été reprises au moins une fois. Les chercheurs en analyse de réseaux sociaux s'intéressent plus particulièrement au réseau de reprises pour établir un autre graphe de relations. (Conover et al., 2011) appliquent leur algorithme de détection de communautés pour extraire différentes catégories d'utilisateurs à partir de ce réseau de relations.

En plus de proposer cette fonctionnalité de reprise de messages, Twitter permet aux utilisateurs de mettre *en favoris* les messages des autres utilisateurs. Cette nouvelle caractéristique donne lieu à plusieurs exploitations, il est possible de la considérer « *tweet à tweet* » ou plus globalement sur l'ensemble des publications d'un utilisateur. Et une fois encore, cette variable pourrait se traiter jour après jour pour voir l'évolution de la popularité de certains contenus sur lesquels on repasse alors qu'ils peuvent parfois être très ancien.

Ces caractéristiques sont plutôt à considérer sur le tard, plus on attend plus il y a de chance d'engranger des reprises et des mises en favoris. Indiquant que malgré le temps qui passe, nos publications continuent à intéresser les autres utilisateurs. Toutefois, comme la reprise est une opération plus facile, plus comprise et plus utilisée que la mise en favoris, la littérature donne plus d'impor-

tance à la reprise qu'à la mise en favoris.

Arrêtons-nous enfin sur les mentions que l'on considère comme entrantes et sortantes, en commençant par parler de ces dernières. Cette caractéristique fait référence à la possibilité de mentionner explicitement un autre utilisateur dans ses messages. Cette caractéristique qui indique la propension d'un utilisateur à directement converser avec les autres utilisateurs peut-être abordée selon plusieurs angles. Les «*robots-spammeurs*» sont par exemple connus pour remplir leurs messages de mentions (Wang, 2010). Plutôt que de compter les mentions, certains chercheurs s'intéressent à leur longueur et la place qu'elles prennent dans les messages. A condition d'avoir accès à l'intégralité des données du réseau social (ou au moins aux tweets en question), une caractéristique intéressante à étudier est le nombre de mentions entrantes, c'est à dire le nombre de fois que l'utilisateur est mentionné par d'autres utilisateurs.

5.2.5 Champ lexical et thèmes abordés

Nous nous intéressons ici au contenu des messages produits par les utilisateurs. Nous avons déjà vu que les messages publiés sur Twitter abordent des thématiques diverses et variées. Les discussions allant d'événements ordinaires à des événements de grande ampleur. Contrairement aux médias traditionnels, les messages ne sont pas catégorisés, ni structurés, ce qui a poussé de nombreuses études comme celle menée par (Abascal-Mena et al., 2015) visant à chercher des communautés lexicales d'utilisateurs, c'est-à-dire des groupes d'utilisateurs parlant de mêmes thématiques.

Le vocabulaire utilisé par les utilisateurs est souvent au centre des intérêts des chercheurs (Weren et al., 2014). Le nombre de mots utilisés reste la caractéristique principale, là encore, comme pour le nombre de messages publiés. Les chercheurs considèrent des variantes temporelles en se concentrant sur des périodes plus spécifiques. La présence de mots que l'utilisateur est le seul à utiliser (Ramírez-de-la Rosa et al., 2014) ou encore les entités nommées (de Choudhury et al., 2012; de Silva et Riloff, 2014) sont autant de critères employés pour catégoriser les utilisateurs entre eux. S'appuyant sur les méthodes analysant la distribution statistique des mots dans les messages, décrites dans les chapitres 2 et 4, nous avons proposé d'apprendre à distinguer différentes catégories utilisateurs selon leur façon d'utiliser les mots (Cossu et al., 2015b).

Il est également possible d'utiliser des *n*-grams pour identifier des *prototypes d'expressions* (Pennacchiotti et Popescu, 2011; Lee et al., 2013; de Silva et Riloff, 2014; Weren et al., 2014) spécifiques à un utilisateur ou à un groupe d'utilisateurs. D'autres équipes se sont plutôt focalisées sur la détection de thématiques de prédilections (Weng et al., 2010; Conover et al., 2011; Aleahmad et al., 2014)

ou encore des expressions (unigrams ou multi-termes) typiques d'un utilisateur (Pennacchiotti et Popescu, 2011; Al Zamal et al., 2012; de Choudhury et al., 2012; Makazhanov et Rafiei, 2013).

Enfin, si certains chercheurs se basent uniquement sur le contenu lexical, beaucoup préfèrent le combiner aux différentes caractéristiques que nous venons de décrire (Rao et al., 2010; Conover et al., 2011; de Silva et Riloff, 2014; Vilares et al., 2014; Weren et al., 2014).

5.2.6 Style éditorial

Le contenu des messages peut également faire l'objet d'une étude non lexicale à partir de caractéristiques issues principalement de la recherche d'information. Parmi les descripteurs les plus utilisés, nous pouvons citer les valeurs de longueurs des mots et messages[7]. Cette caractéristique a d'ailleurs été utilisée par (Laasby, 2014) pour la détection de «*robots*». Ces derniers n'utilisent que quelques mots clés sans tenir compte de la structure grammaticale du message. (de Silva et Riloff, 2014; Weren et al., 2014) entre autres ont de ce fait considérer un indicateur de lisibilité des messages (lexicale et grammaticale) pour différencier les utilisateurs, Par exemple, ceux agissant pour le compte d'une entreprise produisent des messages corrects et évitent les répétitions de caractères comme «*quooooiiii ? ! ! ! !*» qui sont à l'inverse bien plus fréquents chez les adolescents.

Les caractères spéciaux (non-alphanumériques) ont également intéressé les chercheurs comme (Ramírez-de-la Rosa et al., 2014). L'utilisation d'émoticônes et d'acronymes (tel que «*LOL*») est souvent spécifique des «*spammeurs*» qui remplacent les lettres par certains symboles pour passer au travers des filtres anti-spam. Pour d'autres chercheurs (de Silva et Riloff, 2014), la présence d'émoticônes permet d'affiner l'analyse d'humeur. (Rao et al., 2010) sont même encore plus précis en annonçant pouvoir analyser l'humeur selon le sexe après avoir découvert que les femmes sont de plus grandes utilisatrices d'émoticônes que les hommes.

Mots-dièses et liens hyper-textes sont également plébiscités par les chercheurs et plus particulièrement leur nombre et leur variété. Certaines équipes se concentrant sur la taille du lexique de liens et mots-dièses d'autres (Benevenuto et al., 2010; Chu et al., 2012) sur le taux de messages contenant un des deux éléments ou sous différentes formes, leur nombre par message (Pennacchiotti et Popescu, 2011; Uddin et al., 2014). Les utilisateurs souhaitant diffuser de l'information ont généralement tendance à intégrer des liens (Java et al., 2007), tout

7. Parfois en nombre de mots pour ces derniers, plus souvent en nombre de caractères dans les deux cas.

comme les *spammeurs* selon (Benevenuto et al., 2010) ce qui fait que cette caractéristique n'est pas vraiment représentative. Les spammeurs ont d'ailleurs l'habitude de répéter plusieurs fois un même lien et à utiliser des services de réduction de liens (Benevenuto et al., 2010; Wang, 2010) pour cacher le site (souvent malicieux) vers lequel l'internaute est renvoyé (Chu et al., 2012; Ghosh et al., 2012). Toutefois, Twitter propose maintenant de raccourcir automatiquement les liens présents dans les messages via son propre service. Cette nouvelle fonctionnalité impose donc une étape supplémentaire si l'on souhaite bénéficier des résultats des travaux sur le sujet.

Les capitalistes sociaux ne sont pas en reste, en ayant une propension à remplir leurs messages de mots-dièses et liens hyper-textes. Les capitalistes sociaux sont généralement bien identifiés et n'ont pour seule vocation que de remplir leur liste d'abonnés et augmenter leur visibilité (e.g. `#FollowMeBack`, `#FMIFY`, cf. (Dugué et Perez, 2014)). Ce comportement est à l'opposé des utilisateurs veillant à la plus grande diversité entre leurs tweets pour répondre au même critère de maximisation d'audience. Cette pratique a amené (Lee et al., 2011, 2013) à considérer un indicateur d'auto-similarité entre les tweets de ces utilisateurs. Les «*spammeurs*», quant à eux, ont plutôt tendance à répéter plusieurs fois un même contenu ou des tweets très similaires (Lee et al., 2010; Wang, 2010; Lee et al., 2011).

5.2.7 Données externes

Les résultats de moteurs de recherches ou encore les scores et catégories produits par plusieurs entreprises spécialisées dans le profilage représentent toutes les informations que l'on ne peut pas directement obtenir à partir de Twitter. Le nombre de résultats donnés par les principaux moteurs de recherche pour un nom d'utilisateur est un bon indicateur de son niveau d'indexation et *a priori* de son importance sur le réseau. Des entreprises comme Klout (Klout, 2015) et Kred (Kred, 2015) sont apparues sur le marché avec le développement de Twitter sans lien avec le réseau social. Ces dernières sont spécialisées dans le profilage d'utilisateur de réseaux sociaux et plus particulièrement sur l'analyse d'influence (selon leur propre mesure). Le procédé utilisé par Klout est tenu secret alors que celui de Kred est disponible publiquement et se compose de deux scores : «*Influence*» (comment les tweets de l'utilisateur sont perçus par les autres utilisateurs ?) et «*Outreach*» (Dans quelle mesure l'utilisateur contribue à la diffusion des tweets d'autres utilisateurs ?). Toutefois la solution apportée est limitée car l'élaboration de ces scores est souvent tenue secrète, ce qui limite l'interprétation des résultats. D'ailleurs, de nombreux chercheurs ont montré que ces outils pouvaient être trompés par un simple «*bot*» qui apparaissait comme étant influent selon Klout et Kred.

Comme il est possible d'accéder à Twitter depuis plusieurs plate-forme et de publier des messages en utilisant des logiciels prévus à cet effet, certains chercheurs (Chu et al., 2012; Huang et al., 2014; Dugué et al., 2015) se sont intéressés à la manière dont l'utilisateur avait accès au réseau (accès mobile) et à l'application utilisée. On peut évidemment se douter que ceux utilisant de lourdes suites logicielles payantes sont des professionnels agissant pour le compte d'une société ou d'une agence de communication.

5.2.8 Discussions

Nous avons montré dans cette section les différentes caractéristiques considérées dans la littérature pour définir un profil Twitter. S'il y a bien une conclusion importante à cette section, c'est celle-là : nous sommes dépendants du temps dans la sélection et le traitement de toutes les caractéristiques. En effet si contrairement à la majorité des médias sociaux, Twitter permet la collecte de données cela n'est pas sans conditions. Il est difficile d'avoir publiquement et gratuitement accès à plus de 1000 des derniers tweet publiés par un utilisateur. De plus, cet accès est limité à 180 requêtes par tranche de 15 minutes, pour la liste des amis, la collecte se limite à 1000 listes d'abonnés par tranche de 24 heures. Enfin, si l'utilisateur supprime son compte ou ne souhaite plus apparaître publiquement, il devient impossible de récupérer une quelconque information. De plus, il existe des utilisateurs pour lesquels certaines informations ne sont pas récupérables (exemple « *Eminem tweete peu et ne suit personne* » [8]) rendant alors difficile l'application de certaines méthodes.

Ces différentes contraintes se rajoutent à la disponibilité des collections de données, souvent incomplètes ou vite dépassées, ainsi qu'à la reproductibilité de certains résultats, les caractéristiques des utilisateurs considérés ayant largement évolué avec le temps. Dans d'autres cas, par exemple pour les méthodes se basant sur les réseaux d'interactions, c'est le temps de calcul (et celui de récupération des données) qui atteint rapidement des valeurs irréalistes si l'on souhaite s'intéresser aux utilisateurs ayant plusieurs millions de messages ou d'abonnés.

Les méthodes se basant sur les variations dans le temps de certaines caractéristiques comme celles de (Lee et al., 2011) ou mesurant des taux de variation sur de longues périodes ne sont pas reproductibles à moins d'être en mesure de récupérer les données exactes utilisées par les auteurs au moment de leurs travaux.

8. Le chanteur n'a seulement publié que 571 messages, son compte possède environ 20 millions d'abonnés et aucun abonnement voir : https://twitter.com/eminem

Une autre remarque sur ces caractéristique : elles peuvent être considérées sous plusieurs formes et être traitées avec plusieurs méthodes différentes. D'ailleurs, l'ensemble des travaux basés sur les contenus et autres caractéristiques issus des domaines du traitement automatique de la langue et de la recherche d'information sont à prendre avec précautions car elles dépendent des différents prétraitements effectués sur les contenus. Les pré-traitements les plus communément admis dans la littérature, incluant la suppression des symboles non alpha-numériques (ponctuations, émoticônes, liens hyper-textes ainsi que les dièses et arobases mots-dièses et mentions), la « *minusculisation* » des mots et la réduction des caractères répétés (par exemple réduire « *mdrrrrrr* » en « *mdr* »), amènent à considérer des lexiques finalement très différents. Nous faisons l'impasse sur les anti-dictionnaires propres à chaque équipe de recherche et dont l'éventuelle reproductibilité ne fait encore que rarement partie des critères de développement.

Enfin, pour compliquer les choses, certains auteurs définissent de nouvelles caractéristiques à partir de combinaisons de critères plus basiques. Par exemple Tommasel et Godoy (Tommasel et Godoy, 2015) définissent tout un ensemble de ratio d'abonnements et abonnés avec les valeurs de reprises et mentions. Lee *et al*. considèrent plutôt selon plusieurs critères la place occupée par le contenu propre au réseau (« *mention* » et « *mot-dièse* ») sur la place totale occupée par le contenu en langue naturelle réellement intelligible du message (Lee et al., 2011). Cette approche de combinaison est d'ailleurs largement reprise avec d'autres caractéristiques par d'autres équipes (Benevenuto et al., 2010; Rao et al., 2010; Wang, 2010; Chu et al., 2012; Anger et Kittl, 2011; Uddin et al., 2014).

Comme énoncé précédemment, étant donné la quantité de variables et la variété de leurs formes, nous allons par la suite ne nous intéresser qu'à une partie de ces caractéristiques que nous utilisons pour la prédiction de l'influence.

5.3 Expériences

Les données fournies dans le cadre du challenge RepLab 2014 se consacrent directement à la problématique de détection du niveau d'influence d'un grand nombre d'utilisateur au regard d'une thématique donnée. D'ailleurs pour la première fois avec Replab 2014, il est possible de vérifier la pertinence des mécanismes habituellement utilisés pour la détection d'influence sur les réseaux sociaux (influence dite « *en ligne* ») lorsqu'ils sont appliqués à la détection d'influence réelle (dite « *hors ligne* »).

Beaucoup de recherches portant sur l'analyse de l'influence au sein des médias sociaux s'intéressent à de grandes populations d'utilisateurs (souvent l'en-

semble des membres du réseau) sans considération pour les thématiques. Il apparaît pourtant évident que les utilisateurs des médias sociaux réagissent d'une part à des événements précis [9] et d'autre part de manière spécifique à chaque domaine auquel l'événement est rattaché. Face à ce constat, RepLab apporte une valeur ajoutée à la problématique en centrant l'étude sur deux domaines en particulier les secteurs automobiles et bancaires. Même si cela tend à limiter la validité de l'étude au spectre de ces deux domaines cela permet néanmoins de proposer une réponse certes partielle mais bien plus concrète à la question du profilage.

5.3.1 Méthodologie proposée

Nous proposons pour répondre à ce défi de catégorisation d'utiliser les méthodes que nous avons décrites dans le chapitre 2. Nous divisons notre traitement selon les critères du domaine de l'utilisateur et de la langue dans laquelle l'utilisateur s'exprime. Nous proposons une première série d'expériences (que nous appelons «*jointes*» par la suite) qui consiste à traiter l'ensemble des utilisateurs sans distinction de domaine ainsi que leur message sans distinction de la langue dans laquelle ces derniers sont écrits. Dans un deuxième temps, nous proposons justement d'effectuer un traitement en deux étapes en distinguant domaine et langue (approche que nous appelons «*séparée*» par la suite. Les scores de confiance associés à chaque hypothèse sont ensuite combinés pour permettre enfin de constituer le classement ou affecter un utilisateur à sa catégorie. Ces scores correspondent à l'emploi de la similarité cosinus décrite dans la section 2.2.5.

L'approche que nous proposons connaît également deux petites variantes. Dans les deux cas nous estimons une similarité entre un document d et une catégorie C par exemple «*influant*»et «*non-influant*», la nature du document d variant selon l'approche entre un message ou un profil complet. Nous détaillons ces variantes ci-dessous.

La première de ces variantes est appelée *User-as-Document* (UaD) (Kim et al., 2015; Cossu et al., 2014, 2015a). Elle consiste à regrouper l'ensemble des publications d'un utilisateur pour créer un énorme document d qui deviendra le représentant de tout ce qu'écrit l'utilisateur. Nous faisons de même ensuite avec tous les utilisateurs d'une classe pour créer un document C représentant l'ensemble des utilisateurs appartenant à cette classe. Dans le cas de la détection d'influence, le document C représentant l'ensemble des utilisateurs influents (i.e la classe des «*influants*») est en fait la concaténation de tous leurs mes-

9. Il n'est même pas envisageable de pouvoir humainement considérer tous les événements d'une heure sur Twitter.

sages. Lorsque nous appliquons la séparation en langue nous obtenons deux documents, un document C_x pour les influents dans une langue et C_y pour ceux qui s'expriment dans une autre langue. A ce moment là, un utilisateur s'exprimant dans les deux langues peut se retrouver affecté à deux classes différentes, par exemple il est considéré comme influent si l'on considère C_x alors qu'il ne l'est pas si l'on s'intéresse uniquement à C_y. Cela implique d'avoir à considérer une stratégie de combinaison en amont de l'affectation à une classe. Nous avons simplement décidé de pondérer la décision pour chaque langue selon la proportion de messages que l'utilisateur a pu écrire dans chacune d'entre elles. Par exemple, si un utilisateur s'exprime deux fois plus souvent en anglais qu'en espagnol (deux fois plus de tweets), la décision obtenue pour un document construit à partir des messages en anglais aura un poids double par rapport à la décision obtenue avec les données en espagnol.

Nous appelons la deuxième variante *Bag-of-Tweets* (BoT) (Cossu et al., 2014, 2015a). Cette dernière se focalise sur les messages et non les utilisateurs. Cette fois-ci, le document d (dans l'équation cosinus 2.1 du chapitre 2) correspond aux tweets. L'utilisateur est représenté par l'ensemble des messages qu'il a publiés. La classe est, quant à elle, composée de l'ensemble des messages des utilisateurs lui ayant été attribués au moment de l'annotation. Nous estimons ensuite la similarité entre chaque message d'un utilisateur et chaque classe. La classification est obtenue à partir d'une stratégie de décision qui consiste à retenir la catégorie ayant obtenu la majorité des votes selon deux critères : le nombre de votes (règle appelée *Count*) et la somme du poids de ces votes (règle appelée *Sum*). Le classement est obtenu à partir du score donné par la stratégie de décision. Cette stratégie peut s'adapter également au cas où l'utilisateur s'exprime dans plusieurs langues différentes, chacune étant pondérée comme énoncé précédemment.

5.3.2 Evaluation et discussions

Etant donné le nombre limité d'utilisateurs «*influents*» (moins de 30%), les organisateurs ont défini le problème comme étant similaire à celui d'un moteur de recherche qui pour chaque thématique aurait à retourner les utilisateurs les plus pertinents, ici les plus influents. Les organisateurs ont décidé d'évaluer les soumissions des participants à la tâche de détection d'influence en utilisant la MAP , celle-ci en étant en effet bien adaptée à notre application. Cette dernière ayant été décrite dans le chapitre 3 nous n'y reviendrons pas. Les participants au concours ont été comparés selon la MAP obtenue pour chaque thématique ainsi qu'avec la MAP moyenne sur les deux thématiques. Plus récemment, Ramirez *et al.* ont proposé d'évaluer le problème de détection d'influence comme étant une tâche de catégorisation binaire visant à séparer les «*influants*» des

« *non-influants* » (Ramírez-de-la Rosa et al., 2014). Nos propositions seront donc évaluées selon les deux protocoles proposés dans la littérature et décrits dans le chapitre 3.

5.3.3 Classement d'utilisateurs par niveau d'influence : comparaison de performances

Nous proposons de commencer par discuter des résultats de l'évaluation officielle de RepLab. Notons, à titre indicatif, que les organisateurs de RepLab ont proposé un système basique qui classe les utilisateurs par nombre décroissants de suiveurs. Ce système revient à considérer simplement que, plus l'utilisateur est suivi, plus il est influant.

La ligne (Aleahmad et al., 2014) dans le tableau 5.1 correspond au groupe *UTDBRG* (cf. résultats officiels (Amigó et al., 2014)) qui a utilisé l'information des « *tendances du moment* » en supposant que les utilisateurs « *influents* » étaient à la pointe des dernières discussions tendances sur le réseau social (Aleahmad et al., 2014). Selon les résultats publiés, cette méthode est la plus performante en obtenant la meilleure MAP sur le domaine automobile (.721) et la MAP moyenne la plus élevée (.565) parmi tous les participants.

Le groupe *UAMCLYR* a proposé une combinaison de caractéristiques de profil avec des caractéristiques liées au *style éditorial* (richesse du vocabulaire, utilisation des mots et symboles) en utilisant des champs de Markov aléatoires (Villatoro-Tello et al., 2014). Dans une perspective plus axée traitement automatique des langues, les groupes *ORM_UNED* (Lomena et Ostenero, 2014) et *LyS* (Vilares et al., 2014) se sont intéressés aux marqueurs morpho-syntaxique comme caractéristiques additionnelles pouvant être extraites des contenus textuels. La proposition du groupe *LyS* basée sur le sac de mots des descriptions de profil a d'ailleurs obtenu la meilleure MAP pour le domaine banque (.524) et le second score moyen sur l'ensemble des participants (.563).

Lors de notre participation officielle au challenge, nous avons axé notre participation sur l'hypothèse que les « *influants* » utilisaient des termes spécifiques dans leurs messages (Cossu et al., 2014). Chaque utilisateur a donc été modélisé à partir du contenu des messages qu'il a pu publier. Ensuite à partir d'une méthode de « *K plus proches voisins* », chaque utilisateur est rapproché de ceux qui lui ressemblent [10]. Chacun des *K plus proches voisins* votant alors proportionnellement à son niveau de similarité pour l'une des catégories : « *influents* » ou « *non-influents* ».

10. La similarité entre deux utilisateurs est estimée à partir de la représentation UaD de chaque utilisateur.

Nous avons plus tard proposé une variante de notre approche en considérant un critère d'optimisation différent (maximisation de la MAP et non F-Score de catégorisation) et observé de très nettes améliorations de performances obtenant alors les meilleurs résultats observés jusque là sur ces données avec une MAP à .764 pour l'automobile, .652 pour la banque et une moyenne à .708 (Cossu et al., 2015a). Puis à partir d'une méthode basée une similarité cosinus en suivant la même modélisation nous sommes arrivés à obtenir des résultats encore plus élevés, en atteignant une MAP .803 pour le domaine automobile tout en gardant une MAP honorable de .626 pour les banques, la MAP moyenne de .714 étant alors légèrement supérieure aux travaux précédents (Cossu et al., 2015b).

Ce niveau de performance sur le domaine automobile reflète probablement une tendance des « *influents* » à rester en alerte sur les dernières innovations des constructeurs automobiles à l'inverse la contre performance de cette approche avec les banques laisse supposer que l'influence est cette fois dépendante de discussions plus techniques et spécialisées, ce qui peut également expliquer les bonnes performances de la méthode basée sur l'analyse des contenus que nous avons proposée (Cossu et al., 2014). Nous pouvons faire ressortir deux conclusions qui font l'unanimité parmi les participants. La première est que les deux domaines sont bien différents, l'un étant plus difficile que l'autre car les performances observées sont pour tous les participants largement supérieures sur le domaine automobile. La deuxième conclusion est que la fonction témoin proposée par les organisateurs est dépassée par quasiment l'ensemble des participants (MAP .370 et .385).

L'ensemble des résultats obtenus pour chaque domaine, pour la tâche de classement des utilisateurs, se trouve dans le tableau 5.1. La MAP moyenne est utilisée pour classer les différents systèmes. Notons une fois encore que les performances pour les banques sont presque toujours inférieures à celles obtenues pour le domaine automobile.

Dans leurs travaux, (Lomena et Ostenero, 2014; Vilares et al., 2014; Villatoro-Tello et al., 2014) ont pu tirer parti des différentes caractéristiques telles que le *style éditorial* combiné dans un système de SVM pour ordonner les utilisateurs selon leur niveau d'influence. Lors de nos expérimentations, avec des valeurs de caractéristiques récupérées *a posteriori*, nous n'avons à notre grande surprise pas été en mesure de reproduire leurs niveaux de performance, soit par altération des données avec le temps soit car le processus de traitement n'était pas complet. Notre contre performance à ce niveau n'est en soi pas très étonnante. Une mauvaise dénomination de caractéristiques ou un choix différent de variables (valeur moyenne ou écart-type) suffisent à perturber le processus d'apprentissage. Le meilleur résultat que nous avons pu obtenir avec ces approches se trouve juste au dessus de la fonction témoin. D'ailleurs à l'instar de la fonc-

Méthode	Automobile	Banques	Moyenne
UaD Separated	**.803**	.626	**.714**
(Cossu et al., 2015b)	.764	**.652**	.708
BoT Separated Sum	.779	.628	.703
BoT Separated Count	.762	.592	.677
UaD Joint	.735	.538	.636
BoT Joint Sum	.699	.526	.612
BoT Joint Count	.626	.504	.565
(Aleahmad et al., 2014)	.721	.410	.565
Nombre de tweets	.332	.449	.385
Best Regression	.424	.338	.381
RepLab fonction témoin	.370	.385	.378
Klout score	.304	.275	.289

TABLE *5.1* – *Performances sur les données d'évaluation RepLab* 2014 *des approches proposées pour la tâche de classement d'utilisateur par niveau d'influence. Les approches sont triées selon la MAP moyenne obtenue (dernière colonne) et les meilleurs scores sont indiqués en gras.*

tion témoin proposée par les organisateurs basée sur le nombre d'abonnés, la caractéristique du nombre de messages publiés par l'utilisateur permet d'obtenir un classement honorable pour les banques et pourrait se substituer à la fonction témoin dans ce cas précis. Toutes les autres caractéristiques ont montré des résultats bien inférieurs à ceux proposés par la fonction témoin.

Un de nos résultats le plus surprenant est sans conteste la mauvaise performance du classement à partir du score Klout, surprenant car ce dernier a été précisément conçu pour mesurer l'influence à la fois «*en et hors ligne*».

5.3.4 Classification d'utilisateurs par selon leur influence : comparaison de performances

La question de la détection d'influence n'ayant pas été officiellement évaluée sous cet angle, il n'existe pas système fonction témoin auquel il est possible de se comparer. Ramirez *et al*. ne comparent d'ailleurs leurs méthodes à aucune fonction témoin. Cependant, la disproportion des classes influents (31%) versus non influents (69%) dans la collection entraîne de fait une fonction témoin forte et totalement non-informative qui consisterait juste à considérer tous les utilisateurs comme étant «*non influents*». Cette fonction témoin (notée «*MF-fonction témoin*» dans le tableau 5.2) de la catégorie majoritaire obtiendrait un F-Score de .50 et une précision de .69. Ramirez *et al*. ont obtenu un F-Score (macro) de .696 et .693 pour les domaines Automobile et Banque avec une moyenne de .694

alors que Cossu *et al*. n'ont réussi à obtenir que .40 (Cossu et al., 2015b).

Méthode	macro F-Score Automobile	macro F-Score Banques	Moyenne
UaD Separated	**.833**	**.751**	**.792**
(Cossu et al., 2015b)	.812	**.751**	.781
BoT Separated Sum	.817	.709	.763
BoT Separated Count	.786	.702	.744
UaD Joint	.782	.682	.732
(Ramírez-de-la Rosa et al., 2014)	.696	.693	.694
BoT Joint Sum	.725	.641	.683
BoT Joint Count	.725	.641	.683
MF-fonction témoin	.500	.500	.500

TABLE 5.2 – *Performances sur les données d'évaluation RepLab* 2014 *des approches proposées pour la tâche de catégorisation d'utilisateur par niveau d'influence. Les approches sont triées par macro F-Score moyen (dernière colonne) et les meilleurs scores sont indiqués en gras.*

L'utilisation des caractéristiques de manière individuelle (ou bien regrouper par catégorie) avec des SVM ou des régressions logistiques ne s'est pas montrée suffisante pour proposer une séparation pertinente des utilisateurs. Aucune approche n'a pu dépasser le niveau de résultats d'une simple fonction témoin qui consisterait, sachant qu'il y a 70% de «*non-influents*», à considérer que tous les utilisateurs ne sont pas «*influents*». L'utilisation d'un seuil pour maximiser la partition en deux parties de la population d'utilisateur ne s'est pas montrée pertinente non plus. L'approche que nous proposons et ses différentes variantes se comportent efficacement dans cette tâche dépassant de loin les performances de la fonction témoin et celles dites «*état de l'art*». Quel que soit le domaine, ou la tâche (classement cf. tableau 5.1 ou catégorisation cf. tableau 5.2) les deux approches UaD et BoT ont montré la même sensibilité au traitement séparé langues même si il difficile de conclure sur la significativité des différences. L'approche UaD semble toutefois se montrer la plus performante des deux et ce la encore malgré l'absence de significativité (selon un t-test classique). S'agissant de la stratégie de décision (nombre de votre et somme), le nombre de votes s'avère être légèrement moins performant. Tenir compte de la confiance associée à chaque vote (score ou probabilité fournie par le système automatique) semble de toute façon être plus pertinent.

5.4 Conclusions

Nous nous sommes intéressés dans ce chapitre à la problématique de la caractérisation de profils d'utilisateurs et plus particulièrement à la caractérisation de profils des utilisateurs du réseau social Twitter à partir d'une part des caractéristiques utilisées dans la littérature et d'autre part de méthodes simples d'analyse automatique des contenus textuels des messages publiés par ces derniers.

Nous avons commencé par faire le point sur les différentes caractéristiques habituellement utilisées pour répondre à ces questions, caractéristiques pour la plupart issues de domaines de recherche tels que l'analyse de réseaux sociaux, le traitement automatique des langues et la recherche d'information. Nous avons montré qu'il était possible de catégoriser ces caractéristiques selon les différents concepts qu'elles illustrent.

Ensuite, à partir d'expérimentations focalisées sur la question d'identification et classement de personnalités influentes à partir de leur compte Twitter, nous avons mis en avant deux résultats. Tout d'abord, les caractéristiques typiquement utilisées en Analyse de Réseaux Sociaux pour répondre à des tâches de détection de «*spammeurs*», «*capitalistes sociaux*» ou les métriques permettant de détecter les nœuds importants du réseau ne sont pas pertinentes dans notre cas d'étude et d'autant plus particulièrement pour répondre à la question de la détection d'influence telle qu'elle a été définie dans le cadre de la campagne RepLab 2014. Cela est d'autant plus surprenant que l'on pourrait s'attendre à ce que l'influence de ces personnes dans leur domaine se matérialise également au travers des caractéristiques de leur compte Twitter les plus étudiées.

Puis, comme cela a pu être constaté dans la littérature pour d'autres questions de caractérisation de profil d'utilisateurs Twitter, nous avons vu que l'approche consistant à représenter un utilisateur par son vocabulaire (sous la forme d'un sac de mots ou de messages) et plus précisément la manière dont ce vocabulaire est utilisé amenait à des résultats bien plus élevés dépassant même de loin ceux qualifiés *d'état de l'art* [11].

Nous pouvons conclure à partir de nos expériences que d'une part écrire des messages en révèle bien plus sur soi que l'on ne pourrait le croire. D'autre part, la façon d'écrire ces messages relève du domaine dans lequel l'utilisateur est actif et s'exprime. Les mots utilisés et la manière de les utiliser variant d'un domaine à l'autre. Quel que soit le domaine, les «*influants*» semblent avoir un comportement éditorial distinct des autres utilisateurs.

11. La littérature autour de la campagne PAN (sur les questions de détection d'âge, genre et personnalité) avance les mêmes conclusions.

Nous avons également observé que la langue dans laquelle l'utilisateur s'exprime et les traitements plus spécifiques que l'on peut considérer d'une langue à l'autre permettent d'améliorer les performances. Cela indique aussi que les termes issus d'une langue prennent un sens différent lorsqu'ils sont utilisés dans le contexte d'un message dans une langue différente. Enfin bien sûr, même si ne l'avons pas étudiée, la détection d'un métier spécifique et d'un type particulier d'utilisateur pourrait bénéficier d'analyse approfondie de l'information de géo-localisation du profil ou des messages. En effet, les journalistes peuvent «*twitter*» directement depuis le « *point chaud* » à partir duquel ils réalisent leur reportage ou depuis leur bureau à l'instar des déclarations de personnalités respectées du milieu de la finance. Il reste toutefois discutable en terme de respect de la vie privée d'utiliser cette donnée même si celle-ci est donnée volontairement par l'utilisateur lui-même et accessible publiquement.

Il reste important de signaler que la validité de toutes ces conclusions se limite aux données considérées et avec les restrictions supplémentaires des seuls domaines étudiés : banques et constructeurs automobile. Comme toutes les tâches d'apprentissage automatique, la qualité de l'annotation manuelle des données utilisées pour l'entraînement des méthodes automatiques établit également un plafond de performance que peuvent atteindre les systèmes automatiques. Ce plafond tend toutefois à se rapprocher de la performance d'un annotateur humain, preuve que les méthodes d'apprentissage automatique arrivent à reproduire des jugements d'experts sous réserve que ces derniers n'aient pas été assistés par un système automatique. Dans ce cas, comme pour toutes les tâches similaires, développer un algorithme qui s'approche de celui utilisé par l'assistant automatique pour effectuer l'annotation fausserait les résultats et invaliderait la comparaison avec d'autres méthodes.

Plus le temps passe, plus la qualité de l'annotation s'altère, certains utilisateurs ayant clôturé leur compte alors que d'autres ont pris de l'importance (nous avons vu plus haut que la notoriété s'acquiert au fil du temps), ce qui limite de fait la reproductibilité future de l'ensemble des nos expérimentations. Lors de RepLab 2014, les organisateurs (Amigó et al., 2014) n'étaient pas en mesure de statuer sur un élément justifiant significativement un niveau de performance (que ce niveau soit élevé ou fiable). A l'instar de ces derniers, (Rangel et al., 2014, 2015) n'ont pas été non plus en mesure de conclure sur la capacité d'une caractéristique ou d'un algorithme pour répondre aux questions soulevées dans le cadre de PAN. Ces derniers ont néanmoins observé que les méthodes de profilage basées sur l'analyse automatique des contenus textuels se positionnaient parmi les plus efficaces du challenge. Ces observations corroborent les hypothèses que nous avions émises à partir de nos expériences sur l'intérêt de caractériser un utilisateur par ce qu'il dit plus que par ce qu'il est sur le réseau.

Chapitre 6

Visualisation d'information

Sommaire

Dans les chapitres précédents, nous avons présenté différentes méthodes permettant de détecter les contenus porteurs d'opinions et de catégoriser les utilisateurs selon différents critères. Dans ce chapitre, nous décrivons la dernière contribution de cette thèse, qui porte sur la visualisation de l'information. Nous définissons deux axes d'analyse de l'information, le premier est basé sur une méthode de modélisation conceptuelle et la seconde sur une méthode de sélection de l'information pertinente au sein des médias sociaux. Nous proposons dans ce chapitre des approches complémentaires permettant d'une part de générer une visualisation de l'image d'une entité et d'autre part de sélectionner un nombre limité de contenus par rapport à une masse plus importante de messages. Notre objectif est de présenter à l'analyste les contenus les plus représentatifs des réactions de différents types d'utilisateurs par rapport aux événements de sa marque mais également d'indiquer en quoi ces contenus sont représentatifs ou du moins méritent l'attention de l'analyse.

6.1 Introduction

Dans les chapitres précédents, nous avons vu que les différents médias sociaux nous abreuvent d'une quantité astronomique d'informations subjectives revêtant différentes formes (documents textuels dont le contenu et la taille varient fortement) : articles, billets, tweets, etc. . Etant donné le trop grand nombre de messages laissés par les utilisateurs, nous pensons que le défi peut être relevé, si l'on arrive à réduire la masse à un sous-ensemble de documents qui rend possible l'analyse. Pour cela, il faut savoir comment les ordonner et retenir ceux qui sont les plus «*pertinents*». Nous proposons une chaîne de traitement automatique pouvant être facilement et rapidement appliquée aux problèmes de sélection des documents (ici des tweets émis par un utilisateur donné ou relevant d'un concept particulier) en utilisant des systèmes de résumé automatique de documents et des méthodes qui s'apparentent à celles employées en RI. Les méthodes proposées permettent de détecter *a posteriori* les événements ayant suscité l'intérêt des utilisateurs d'un média social.

Beaucoup de services utilisent des fonctionnalités basées sur des mots-clés pour proposer des services d'analyses d'images. Twitter par exemple, offre directement cette possibilité via le «*Trends service*» pour rester au fait de l'humeur du marché et des derniers événements. Si la détection a fait l'objet de nombreuses recherches et semble être une problématique en bonne voie de résolution, l'histoire de ne s'arrête pas là. En effet, ce genre de service ne donne accès qu'aux 1500 derniers messages contenant ces mots-clés que ces derniers soient informatifs ou non. De plus le choix de ces mots-clés devient déterminant quant à l'efficacité de l'analyse menée ensuite : entre une masse indigeste d'in-

formations et de trop rares documents. Dans le premier cas d'ailleurs, utiliser des méthodes de résumés automatique générique ne permettrait pas d'obtenir un extrait de textes beaucoup plus informatif. L'analyste se retrouve alors devant des bribes d'indices de l'existence d'événements sans véritablement savoir de quoi il en retourne et probablement pire encore, il n'aurait pas la possibilité de connaître l'opinion des utilisateurs au sujet de ces événements. Ces différentes vues offertes par extraction et mots clés sont d'autant plus sensibles si l'on considère des entités ou concepts ambigus. Les méthodes statistiques décrites précédemment permettent de proposer un classement des documents vis-à-vis de chaque critère d'interrogation des données (concept, thématique, requête spécifique, ...) ainsi que d'extraire les termes les plus représentatifs. Ces derniers permettent d'argumenter auprès de l'analyste les hypothèses du système.

Nous nous proposons d'étudier l'utilisation de méthodes statistiques simples dans des scénarios contrôlés afin d'offrir une alternative aux approches proposées précédemment combinant l'opinion (ou l'alerte) aux axes d'analyse suggérés par des experts. Le scénario se base sur la recherche et l'organisation de liens entre des signaux variés (les systèmes automatiques de catégorisation). Nous pourrions très bien remplacer ces systèmes par de multiples experts en communications ou conseillers, chacun indiquant une opinion ou un point d'action différent. Notre méthode interviendrait ici comme un arbitre ou «*méta-expert*» qui serait là pour agréger et attribuer des priorités aux recommandations de chaque expert.

C'est au travers de ces objectifs que nous proposons ensuite des résumés automatiques guidés en utilisant les différents événements comme des requêtes afin de produire des extraits du flux de messages. Nous considérons comme requêtes des sujets de discussions, des thèmes plus ou moins concrets (le nom d'une affaire, d'un produit ou le concept d'éthique), des étiquettes de *groupe* ou même un tweet donné en exemple pour illustrer un événement. Nous avons ici un double objectif, d'un part réduire le flux de messages à traiter, d'autre part permettre à l'analyste de rapidement assimiler l'essentiel de l'information autour d'un événement et ainsi lui éviter le délicat choix des mots-clés.

6.2 Travaux connexes

Il existe dans la littérature deux grandes lignes associées à la visualisation d'information dans le cadre d'une analyse d'image de marque sur les réseaux sociaux. La première se concentre sur une modélisation indirecte : c'est à dire l'annotation de contenus puis un découpage des données (la granularité étant variable selon le cas d'étude) afin d'établir le plus souvent de groupes de don-

nées autour de différents critères (opinions ou thématiques) dans le but d'alimenter des tableaux de bords. La deuxième, elle, se base sur une analyse plus fine des contenus afin d'extraire des éléments représentatifs des tendances du moment : le dernier événement à la mode, le message le plus représentatif d'une tendance, les termes les plus utilisés dans cette discussion, etc..

6.2.1 Suivi de réputation

Essayer de modéliser une information conceptuelle aux multiples facettes est loin d'être un problème de recherche récent, (Tanaka, 1993) proposait un état de lieux des méthodes qui existaient à ce sujet au début des années 1990. Un peu plus tard, dans les années 2000, le besoin de visualiser l'information à donner naissance à l'expression de « *fouille visuelle de données* » (Chen, 2006). La visualisation de graphes s'est ainsi imposée comme une nouvelle méthode de fouille de données parallèlement aux approches numériques plus classiques. Toutefois, outre le fait qu'il n'existe pas de méthode de visualisation universelle il faut également palier à l'absence de méthodes adaptées au Web 2.0. Car si nous sommes en mesures de construire des visualisations à partir des données à notre disposition, il faut pouvoir être en mesure de savoir réellement ce que représentent ces « *images extraites* » dont nous souhaitons proposer une analyse visuelle.

L'introduction de systèmes automatiques permettant d'analyser l'image de marque d'entités publiques dans le monde de la recherche est assez récente et le développement des campagnes RepLab [1] (Amigó et al., 2013a, 2014), TASS [2] (Villena Román et al., 2013) et ou encore Imagiweb [3] (Velcin et al., 2014) est sûrement le meilleur signe de cet engouement. Ces récents développement ont redéfini ce qu'un expert attend aujourd'hui d'un système automatique. Cet assistant ou système automatique d'analyse de réputation doit être à même d'assister l'analyste en lui suggérant, parmi un flux important de données, des contenus sur lesquels il devrait porter son attention tout en lui indiquant pourquoi ces derniers sont dignes d'intérêt. Nous pouvons considérer qu'il existe trois étapes importantes :

— la sélection des contenus à analyser (étape communément appelée filtrage en recherche d'information) ;
— le traitement des contenus afin de détecter l'opinion selon deux critères : polarité et la thématique (analyse d'opinions, de sentiments) ;
— enfin, la visualisation dans le but de mettre en avant le contenu, tout en proposant de naviguer dans les éléments qui justifient cette mise en

1. http://www.limosine-project.eu/events/replab2013
2. http://www.daedalus.es/TASS2013/corpus.php
3. http://mediamining.univ-lyon2.fr/velcin/imagiweb/

avant.

Toutes ces étapes, et leur évaluation avec des collections de données issues de cas pratiques, font encore aujourd'hui l'objet de toute l'attention de nombreuses équipes de recherche comme nous l'avons vu dans le chapitre 2. C'est notamment sur le filtrage (considéré comme problème de détection et réduction de «*bruit*») que les travaux en recherche d'information se sont illustrés et ont été évalués notamment dans le cadre des conférences TREC. D'ailleurs dans la lignée de ces conférences, d'un point de vue recherche d'information, la modélisation de thématique repose sur des modèles non supervisés [4] comme le modèle TwitterLDA proposé par (Zhao et al., 2011). De notre point de vue, définir correctement ces concepts est l'élément clé pour s'assurer plus tard de leur bonne exploitation. Cette tâche est généralement attribuée à des experts du domaine (chercheurs en sciences politiques, sociologues, analystes de marchés, etc.). Ces experts analysent les contenus, suivent l'évolution de l'opinion et essayent de proposer des taxonomies de thématiques sur lesquelles sont exprimées les opinions. Ces taxonomies ou listes d'aspects permettent une analyse plus fine des opinions et doivent aider à mieux comprendre les évolutions d'opinions. L'ambiguïté et la brièveté des tweets rendent inutilisable la plupart des lexiques d'aspects et de sentiments. (Peleja et al., 2014) ont récemment proposé d'extraire automatique un lexique propre à l'analyse de réputation à partir des données en utilisant une méthode basée sur la LDA.

Le problème est alors de trouver un compromis entre les étapes d'annotation et catégorisation qui tiennent compte d'une taxonomie définie manuellement par des experts comme par exemple les standards donnés par le «*Reputation Institute's Reptrak framework*» [5] dans le cadre de RepLab, ou bien avec les thématiques définies par les politologues du CEPEL pour le projet Imagiweb. Ce problème de prime abord secondaire se révèle être le véritable goulot d'étranglement si l'on souhaite tirer parti de la variété des concepts attachés à la réputation et pourrait presque être considéré comme l'enjeu principal.

6.2.2 Résumé automatique : le cas du Micro-Blog

Les méthodes de résumé automatique de documents textuels peuvent être classées en deux grandes catégories selon le type de résumé qu'elles produisent : le résumé par extraction et le résumé par abstraction. Dans le premier cas, le document source est divisé en phrases. Un score est calculé pour chacune d'entre elles et seules sont conservées pour le résumé final celles qui ont obtenu les scores les plus élevés. L'inconvénient majeur est que le résumé

4. Sans avoir défini au préalable les concepts (thématiques) que l'on souhaite étudier.
5. http://goo.gl/LzubRq

est finalement composé des phrases composant le ou les documents d'origine. Dans le deuxième cas, l'objectif est de produire une représentation abrégée du contenu comme le propose (Torres-Moreno, 2011a). Toutefois, selon la littérature (Boudin, 2008; Villegas, 2013), les outils informatiques n'ont pas encore la maturité nécessaire pour résoudre ce défi ambitieux. C'est la raison pour laquelle le résumé par extraction est toujours aujourd'hui l'approche privilégiée (Torres-Moreno, 2011a).

Les chercheurs ont concentré leurs efforts sur la génération de résumés génériques (Abracos et Lopes, 1997; Teufel et Moens, 1997; Hovy et Lin, 1999). En recherche d'information, les requêtes des utilisateurs (les analystes, managers de marques) sont utilisées pour retourner des portions de documents qui correspondent aux points de l'entité que l'on souhaite analyser. Les travaux de (Tombros et al., 1998; Schlesinger et al., 2001) et (Kupiec et al., 1995) proposent d'utiliser des méthodes plus complexes pour apprendre à sélectionner des phrases en fonction de leur position dans le texte et de différentes métriques associées aux termes. Les méthodes utilisées sont majoritairement statistiques (da Cunha et al., 2007) et basées sur le modèle vectoriel (Salton, 1971; Salton et McGill, 1983) très répandu en extraction d'information. Ces méthodes rendent possible la reproductibilité des résultats dans d'autres langues avec le même niveau de résultats. Les travaux de (Sparck-Jones, 2007; Das et Martins, 2007; Ježek et Steinberger, 2008) établissent un panorama assez exhaustif de l'état de l'art du résumé automatique.

Les médias sociaux, et plus particulièrement Twitter, sont relativement récents et bien qu'ayant déjà fait l'objet de nombreuses recherches en détection et résumé d'événements [6]. Les travaux de (Sharifi et al., 2010) sur la sélection de messages représentatifs d'un événement puis ceux (Zubiaga et al., 2012) sur le résumé de match de football sont devenus les références incontournables auxquelles il est utile de se comparer dès lors que l'on s'intéresse à la problématique du résumé d'événements dans les messages courts. A partir des mêmes données, Mackie *et al.* ont proposé différentes méthodes de résumé ainsi que différents protocoles d'évaluation permettant de comparer les résumés produits par ces méthodes (Mackie et al., 2014a,b). Toutefois, ces évaluations étaient centrées sur l'aspect détection d'événements. Leur objectif étant toujours de trouver, dans une masse de messages sur une courte période, un message considéré comme représentatif d'un événement donné. Pourtant si les performances de ces méthodes sont satisfaisantes et permettent aisément de détecter les principaux événements durant une rencontre entre deux équipes (Zubiaga et al.,

6. Cette tâche est toutefois ramené à la simple détection d'un message qui serait le plus représentatif d'un événement. Cette plate-forme reste finalement assez peu étudiée par les chercheurs en résumé automatique du texte dans l'optique de produire une synthèse des discussions autour du dit événement.

2012), produire le résumé de l'efficacité de la défense d'une équipe durant le match reste un sujet ouvert et peu exploré alors qu'il s'agit pourtant d'un enjeu majeur pour la compréhension des discussions sur les réseaux sociaux. A ce titre, nous pouvons noter l'approche de (Wen et Marshall, 2014), basée sur les modèles de Markov cachés, permettant de classer selon leur importance les messages associés aux *discussions du moment* sur Twitter.

6.3 Méthode de sélection de l'information pertinente

6.3.1 Problématiques

Le résumé de Micro-Blogs est un problème de résumé automatique de textes qui peut être considéré selon deux axes :

— le résumé d'un long document (l'ensemble des messages ayant trait à une thématique) où l'objectif est de déterminer les portions les plus importantes de ce document (c'est à dire sélectionner les messages les plus représentatifs) ;

— le résumé de plusieurs petits documents parlant d'une même thématique, cette fois l'objectif est de compresser cet ensemble ou d'en extraire de l'information.

En s'inspirant des méthodes issues de la recherche d'information, ce résumé peut être contraint par la thématique (ou requête de l'analyste), il est alors question de résumé guidé ou personnalisé. Les éléments ou requêtes que nous considérons pour guider le processus de résumé peuvent parfois être similaires aux événements utilisés par (Zubiaga et al., 2012) avec toutefois deux nuances. Premièrement, la requête est formulée par l'utilisateur du système d'analyse de réputation (ou l'analyste ou l'expert). Cela implique qu'elle peut ne pas être pertinente vis-à-vis des données considérées. Il se peut très bien qu'aucun message n'ait été publié sur un sujet bien spécifique ou qu'à l'inverse le sujet soit trop générique et concerne l'ensemble des messages. Deuxièmement, rien n'empêche un message de répondre à deux problématiques différentes. Ce deuxième cas est le plus problématique du point de l'évaluation car dans la plupart des collections de données annotées, l'annotation ne tient pas compte de ce cas de figure.

6.3.2 Contributions

Nous considérons le problème de résumé de Micro-Blogs comme un problème de sélection de messages. Par exemple, nous souhaitons savoir de quoi un utilisateur parle (ses sujets de prédilection) à partir de sa masse de publications ou bien apporter la réponse à une question longue et complexe comme *l'efficacité de la défense d'une équipe durant le match* ou *le niveau d'innovation des produits proposés par une entreprise* à partir d'un flux de messages. Afin de permettre à l'analyste d'être au fait des sujets d'actualité mais aussi de savoir ce que les utilisateurs en disent.

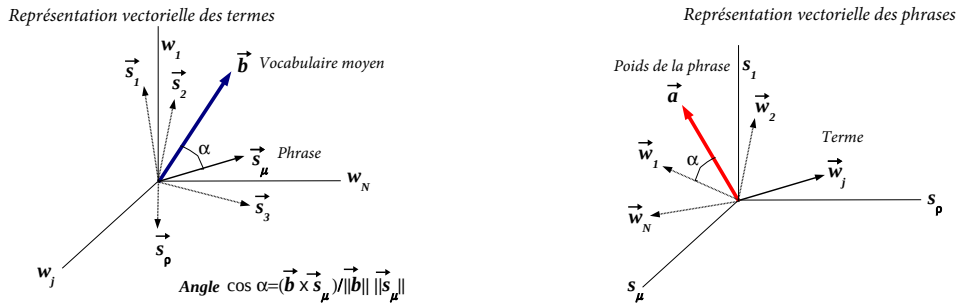


FIGURE 6.1 – *Représentation vectorielle utilisée par le résumeur* ARTEX

$$\omega(\vec{s}) = \left(\vec{s} \times \vec{b}\right) \times \vec{a} \tag{6.1}$$

Nous proposons d'utiliser le système de résumé automatique « ARTEX »[7] décrit par (Torres-Moreno, 2012). Comme le montre la figure 6.1, ce système calcule un score de pertinence $\omega(\vec{s})$ (cf. formule (6.1)) pour chaque phrase à partir du produit vectoriel entre (i) la représentation vectorielle de la phrase considérée, (ii) celle d'une « *phrase moyenne* » censée représenter « *l'ensemble* » des phrases du document et (iii) un vecteur représentant le vocabulaire « *moyen* » du document.

Si on considère $\vec{s}_\mu = (s_{\mu,1}, s_{\mu,2}, \ldots, s_{\mu,N})$ comme étant le vecteur des phrases $\mu = 1, 2, \ldots, \rho$. Le *vocabulaire moyen* $\vec{a} = [a_\mu]$, a été défini comme le nombre moyen d'occurrences du terme N dans la phrase $\vec{s}_\mu$ (voir (6.2)) :

$$a_\mu = \frac{1}{N} \sum_j s_{\mu,j} \tag{6.2}$$

Le vecteur représentant la « *phrase moyenne* » $\vec{b} = [b_j]$ est le nombre moyen d'occurrences de chaque mot j utilisé pour la phrase ρ comme indiqué dans

7. En français : Autre Résumeur de TEXtes

(6.3).

$$b_j = \frac{1}{\rho} \sum_{\mu} s_{\mu,j} \tag{6.3}$$

Le score de pertinence indique donc une centralité de la phrase dans le vocabulaire considéré. Le résumé est ensuite généré en ne conservant que les N phrases ayant obtenu le score le plus élevé (Torres-Moreno, 2014). Ce processus ne comporte aucune phase d'apprentissage ce qui permet de l'appliquer à n'importe quelle collection de documents.

6.4 Modélisation de réputation

6.4.1 Problématiques

Les experts nous proposent un axe d'analyse (parfois plusieurs axes hiérarchisés) qui représente leur vision du problème et n'est de fait pas spécifiquement adapté aux données. En plus de leur proposer des annotations suivant cet axe, nous souhaitons donner aux experts des éléments d'explication concernant à la fois nos annotations et l'axe d'analyse. (Tanaka, 1993) indique qu'il est également nécessaire s'intéresser à la formalisation du problème en accord avec l'analyste qui sera confronté au quotidien à ce que nos outils lui proposeront.

Pour ce faire, nous avons besoin d'une méthode permettant de suivre et visualiser les différents concepts que l'expert souhaite analyser à partir de données concernant l'image de sa marque. En effet, affiner l'axe d'analyse permet indirectement d'améliorer la qualité des méthodes automatiques utilisées pour l'annotation d'opinions. En affinant l'axe d'analyse pour chaque thématique ou aspect de l'entité, puis, en intégrant ces axes affinés dans le processus de traitement des méthodes automatiques d'annotation, nous serons en mesure de fournir à l'expert des points d'actions spécifiques à chacune des thématiques afin de lui permettre d'améliorer l'image de sa marque et sa perception sur ces aspects particuliers.

Nous travaillons à partir d'un petit ensemble des messages annotés par des experts selon plusieurs critères comme la priorité de l'analyste (Mather et Sutherland, 2011) et l'opinion (thématique et polarité) au sujet de l'entité telle qu'est perçue. Même si nous ne pouvons pas nous assurer de la qualité de l'échantillon de messages à notre disposition (Morstatter et al., 2013), nous utilisons cet échantillon pour entraîner une sélection de K systèmes de catégorisation qui permettront d'évaluer l'intérêt d'inclure de nouveaux messages à ceux auxquels l'analyste s'intéresse déjà. Nous appliquons une procédure «*Learn-To-Rank*» pour ordonner, à partir de leurs contenus textuels, des masses non

annotées de messages selon les différents critères (les thématiques et l'opinion). Ces classements nous permettent d'extraire un premier niveau d'interactions entre les critères et puis classer ces interactions thématiques selon leur impact sur l'opinion (polarité) ou la priorité.

Ce travail est généralement effectué manuellement, en prenant le niveau d'importance (priorité) comme principal critère de tri pour ordonner les messages. Il suffit ensuite de descendre ce classement (du message qui semble le plus important à ceux qui ont une importance moindre) tout en considérant les opinions (ou les thématiques) associées aux N premiers messages pour voir celles qui semblent le mieux répondre à la priorité de l'analyste. Dans le cas d'une marque (ou entité) définie par un grand nombre de thématiques ou mentionnée dans un grand nombre de messages voire d'un cabinet de communication gérant plusieurs marques, le besoin d'une méthode pouvant automatiser ce traitement est impératif.

6.4.2 Contributions

La méthode que nous proposons se base sur l'algorithme PLS [8]. Cet algorithme est connu (scientifiquement et professionnellement [9] (Jakobowicz, 2007)) pour l'efficacité de son processus itératif qui alterne avec une fonction d'optimisation par moindres carrés mais aussi et surtout pour sa capacité à pouvoir manipuler toutes sortes de variables et pour représenter une manière inhabituelle de combiner différents systèmes. La méthode PLS permet d'estimer un modélisation dite « *douce* » d'équations structurelles (Wold, 1982) sans pour autant présenter de fonction à optimiser dans le cas général [10].

La modélisation PLS-PM permet de représenter un ensemble de variables comme étant une structure de blocs de variables manifestes (observées). Un document (message ou utilisateur) est donc représenté par un bloc de variables manifestes (par exemple les différents scores de systèmes automatiques pour ce document). Chaque bloc résumant une variable latente qui dépend de toutes les variables manifestes composant le bloc (différentes probabilités d'appartenance à une catégorie pour un message, différentes valeurs de caractéristiques pour un utilisateur). L'algorithme PLS-PM estime alors la meilleure pondération d'une part entre les variables manifestes et latentes puis d'autre part entre les variables latentes et la variable prédite en calculant la solution sous-jacente

8. Nous utilisons l'algorithme Partial Least Squares Path Modeling (PLS-PM) sous son implémentation R via la bibliothèque dédiée http://goo.gl/Di1is0

9. http://www.stat4decision.com/en/

10. La littérature se concentre alors pour les cas particuliers ou justement il existe une fonction à optimiser

du modèle PLS (Henseler, 2010). D'autres détails ainsi qu'une revue complète sur les applications de la méthode sont disponibles au travers des travaux de (Tenenhaus et al., 2004).

Cette approche est utilisée pour vérifier la validité d'un modèle. Elle est dite «*confirmatoire*», ce qui veut dire qu'elle nécessite un modèle conceptuel issu de connaissances d'experts. Toutefois, contrairement à ce que propose la littérature à ce sujet (Jakobowicz, 2007), notre démarche se distingue en ne reposant pas sur modèles préconçus mais en testant de manière expérimentale ce qui émerge des données. Notre objectif est ici d'extraire de l'information à partir des données grâce au modèle lorsqu'un élément n'a justement pas été pris en compte par les experts. De plus le modèle conceptuel permet de bénéficier d'une visualisation claire des relations entre les variables manifestes et latentes. Nous pensons que cette information est précieuse pour les analystes fussent-ils non spécialistes.

6.5 Données et évaluations

Dans cette section, nous décrivons notre protocole expérimental. Cela comprend les données que nous utilisons et les méthodes d'évaluations choisies. Notre objectif est de mesurer la capacité des systèmes automatiques que nous proposons à répondre aux questions suivantes pour les systèmes de sélection d'information :

1. Pour une requête donnée, les systèmes de résumé automatique sont-ils capables de retrouver des messages sélectionnés manuellement pour leur pertinence ?
2. Les messages retournés sont-ils informatifs vis-à-vis de la requête et de l'ensemble des autres messages ?
3. Est-ce possible de profiler efficacement un utilisateur à partir d'un échantillon des messages qu'il a produit ?

Cette modélisation, doit permettre d'extraire des relations entre les divers critères d'analyse de l'image afin d'obtenir une meilleure compréhension de celle-ci.

6.5.1 Collections de données

Nous utilisons pour nos expériences les collections RepLab'2013 et 2014 en partie décrites dans le chapitre 3.

RepLab 2013

Nous utilisons la collection RepLab de l'année 2013 pour proposer un bilan de la réputation de chaque entité point par point. Nous nous focalisons sur l'annotation associée à la tâche de dite de « *clustering* » (appelée « *Topic Detection* »). Pour celle-ci, les participants au concours devaient recréer des paquets de messages traitant d'une même thématique comme ont pu le faire les annotateurs du cabinet Llorente & Cuenca durant l'annotation des données de la campagne RepLab (cf. chapitre 3). Les étiquettes de ces groupes ou thèmes sont des plus diverses : événements, conversations, chansons, commentaires de produits, nouvelles, etc. Nous avons de plus constaté que le nombre de messages par paquet est variable. Pour certains thèmes, nous n'avons qu'un seul message alors qu'il existe des paquets de plus de 20 messages. Nous avons d'ailleurs retiré ces cas particuliers (paquets de moins de deux ou plus de 20 messages) de nos expériences ce qui ramène le nombre de paquets à 3,521. Nous avons toutefois conservé tous les paquets ayant des étiquettes très similaires comme « *photos publiées sur les réseaux sociaux* » et « *photos partagées sur les réseaux sociaux* ». Comme évoqué précédemment, ces cas ont un impact négatif sur l'évaluation, les messages pouvant appartenir à l'un comme à l'autre paquet. L'autre difficulté vient de la complexité des étiquettes. Lorsque le paquet est mono message, l'étiquette est le contenu complet du message. Nous considérons pour la suite de ce chapitre que ces étiquettes seront nos requêtes servant à guider les résumés.

RepLab 2014

Nous avons vu dans le chapitre 3 que les organisateurs de RepLab 2014 fournissaient 600 messages pour chaque utilisateur de la collection. Nous avons considéré dans le chapitre précédent l'ensemble de ces messages associés à chaque utilisateur. Nous proposons cette fois une seconde stratégie (notée « ARTEX » dans les tableaux 6.5 et 6.6) qui consiste à sélectionner un échantillon des 10% de messages les plus « *informatifs* » parmi l'ensemble des publications d'un utilisateur. La sélection est opérée à partir du résumeur automatique « ARTEX » (Torres-Moreno, 2012) que nous avons décrit plus haut. Nous souhaitons vérifier la stabilité des résultats des méthodes que nous avons proposées lorsque celles-ci ont moins de données à leur disposition pour déterminer leurs hypothèses.

Pour la modélisation, afin de proposer une modélisation relativement fiable, nous nous limitons aux 4,720 utilisateurs (respectivement 2,310 pour le domaine automobile et 2,410 pour les banques) pour lesquels nous avons pu col-

lecter l'ensemble des informations [11].

6.5.2 Évaluations

Évaluation de la sélection d'information

L'évaluation des résumés automatiques est reconnue comme étant un défi difficile dans la littérature, comme l'ont montré les travaux de (Mackie et al., 2014a,b) sur le sujet. Nous nous intéressons ici à l'évaluation de la reproduction des «*classes*» définis par les spécialistes de Llorente & Cuenca. Ces «*classes*» sont comme nous venons de le voir (et comme nous l'avions vu dans le chapitre 4) des groupes de messages partageant un même thème. De ce fait, ils représentent une vision précise qu'ont les utilisateurs de l'entité concernée sur ces thèmes particuliers.

Nous proposons une double évaluation, d'abord en évaluant pour chaque requête le niveau de pertinence des messages que les systèmes positionnent en haut du classement en utilisant la MAP décrite dans le chapitre 3. Ensuite, nous extrayons à partir du classement les N messages les plus «*pertinents*» (aux yeux du système) pour constituer un résumé dont nous évaluons la qualité «*informative*» à partir du *cluster* et de l'ensemble des messages en utilisant FRESA (Torres-Moreno et al., 2010). FRESA [12] («*A FRamework for Evaluating Summaries Automatically*») est une méthode d'évaluation de résumés qui permet de s'affranchir des résumés de références (constitués humainement) en calculant une divergence de Kullback-Leibler modifiée $\mathcal{D}(P||Q)$ (Torres-Moreno, 2011b) moyennée sur des ensembles de n-grames de ultrastems (combinaisons d'unigrames, bigrames et SU4-grames) entre un texte source P et un résumé Q. Cette méthode se base sur les travaux de (Lin et al., 2006; Louis et Nenkova, 2009).

Notre objectif est ici de vérifier la capacité des systèmes de résumés automatique à sélectionner efficacement un jeu réduit de messages qui seraient informatifs au regard d'une demande particulière d'un utilisateur.

Évaluation de la modélisation

Lorsque l'on évoque l'évaluation de modèles se pose la question de la représentativité de ces derniers. Toutefois cette représentativité est difficile à es-

11. C'est à dire toutes les caractéristiques de profil définies dans le chapitre 5 ainsi que l'estimation du niveau d'influence par des méthodes automatiques basées sur l'analyse de contenu vues dans le chapitre 4.
12. FRESA est disponible à `http://fresa.talne.eu/`

timer objectivement. Il n'existe à ce jour aucun protocole d'évaluation permettant l'évaluation automatique de la qualité de modélisations. La seule évaluation « *fiable* » connue à ce jour est la confrontation directe du modèle à l'expert. De ce fait, nous nous limiterons à une utilisation descriptive (ou analytique) des modélisations que nous proposons. Les veilleurs préfèrent généralement des interfaces de visualisation traditionnelle (de type graphique temporel, à partir des quels on vérifie des corrélation avec des clichés instantanés de la réalité). Nous proposons de leur apporter une nouvelle visualisation qui présente l'avantage d'être acceptable pour un utilisateur non spécialisé et une représentation type, telles que celles qui existent dans d'autres domaines.

Nous avons vu en section 6.4, la littérature concentre l'évaluation des modèles PLS sur l'optimisation d'une fonction qui n'existe pourtant que dans certains cas particuliers (Jakobowicz, 2007). Ce besoin du vielleur peut finalement se rapprocher des analyses de satisfaction de clientèles comme le définissaient (Fornell, 1992) et (Jakobowicz, 2007). Toutefois, nous nous plaçons dans le cas complexe où le « *processus itératif* » et « *l'alternance entre deux optimisations par moindres carrés* » ne présentent pas de fonction globale à optimiser car nous allons justement considérer tous les modèles possibles et tester expérimentalement leur convergence. Nous utilisons l'index R^2 pour évaluer la qualité du modèle (en maximisant le carré de la somme des corrélations à l'intérieur et entre les variables latentes) en complément de la pertinence que « *observable* » du modèle. Nous nous intéressons donc ici à une *pseudo-évaluation* subjective de la pertinence de la modélisation *vis-à-vis* de notre connaissances de données, nous pensons également à ce que notre modélisation peut apporter à l'analyste, en quoi facilite-t-elle la compréhension du problème ?

6.6 Expériences

6.6.1 Sélection de messages

Nous comparons les performances de la méthode que nous proposons avec le système de résumé automatique « *Cortex* » (Torres-Moreno et al., 2005). Ce dernier se base sur un algorithme de décision multicritères combinant les métriques décrites par (Torres-Moreno et al., 2005) pour évaluer la pertinence d'une phrase (ou d'un document) à partir de sa représentation vectorielle. Enfin, nous avons établi une fonction témoin qui consiste à sélectionner aléatoirement 15 documents et à leur attribuer de manière aléatoire un score de pertinence compris entre 0 et 1. Les systèmes sont évalués sur les 3,521 paquets retenus.

Le tableau 6.1 montre l'informativité moyenne mesurée avec FRESA entre d'une part les résumés proposés par les systèmes automatiques et d'autre part les résumés de discussions constitués manuellement par les annotateurs avec l'assitance d'ORMA comme nous l'avons vu en section 3.3.2.

Résumeurs		
CORTEX	ARTEX	FONCTION TÉMOIN
0.0594	**0.1458**	0.0483

TABLE 6.1 – *Résultats de l'évaluation FRESA entre les sorties des résumés automatiques et les résumés de référence générés manuellement (moyennée sur l'ensemble des requêtes).*

En complément, afin d'évaluer la pertinence des messages sélectionnés par les systèmes de résumé automatique, nous calculons la MAP. Cette évaluation permet de vérifier à quel point la sélection automatique des messages correspond à celle effectuée manuellement par les annotateurs. Cette évaluation est présentée dans le tableau 6.2, les performances sont exprimées en MAP (moyenne sur l'ensemble des résumés) pour chaque système automatique de résumé.

Résumeurs		
CORTEX	ARTEX	FONCTION TÉMOIN
0.0040	**0.1155**	0.0008

TABLE 6.2 – *Résultats de l'évaluation de pertinence selon la MAP entre le classement des phrases produit par les résumeurs automatiques et le résumé de référence (moyennée sur l'ensemble des requêtes).*

Dans ce cas, la fonction témoin qui consiste à sélectionner aléatoirement des messages, pour leur affecter un score lui aussi aléatoire, amène logiquement à un mauvais résultat puisqu'elle n'a quasiment aucune chance de donner les meilleurs scores aux messages attendus, c'est à dire ceux sélectionnés par les annotateurs.

Le tableau 6.3 montre l'informativité moyenne mesurée avec FRESA entre d'une part les résumés proposés par les systèmes automatiques et d'autre part l'ensemble des messages associés à une entité.

Résumeurs			
Reference	CORTEX	ARTEX	FONCTION TÉMOIN
0.0022	**0.0086**	0.0066	0.0061

TABLE 6.3 – *Résultats de l'évaluation FRESA entre les sorties des résumés automatiques et l'ensemble des messages associés à une entité (moyennée sur l'ensemble des requêtes).*

La manière dont les données ont été annotées permet d'expliquer les résultats ci-dessus, intéressants quoique très faibles. En effet, les annotateurs ont regroupé entre eux les messages relevant d'une même conversation ou d'un événement très spécifique, très souvent sans aucun rapport avec les autres discussions. A l'inverse le résumeur automatique essaye de préserver de l'information présente dans l'ensemble des messages même s'il a bien pour objectif premier de retrouver les messages issus d'une même discussion. Cependant, il n'existe pour l'instant aucune piste permettant d'optimiser les deux critères à la fois. Dans ce contexte de conversations multiples entre internautes, c'est sûrement ce critère d'information présente dans l'ensemble des messages qu'il serait utilise d'optimiser peut-être en travaillant sous la contrainte de l'ensemble des résumés à produire à partir de ces données.

Considérons maintenant une requête pour laquelle l'ensemble des systèmes automatiques (y compris la fonction témoin) ont été en mesure de produire un résumé étant un minimum cohérent et intelligible (c'est à dire que les systèmes ont sélectionné au moins un message que les annotateurs avaient déclaré être pertinent) : « *Annie Le's Family Sues Yale* ». Cette requête fait référence au meurtre d'une étudiante sur le campus de l'université de Yale. Les systèmes automatiques ont obtenu une MAP de .3571 pour « *Artex* », .0714 pour « *Cortex* » et .0179 pour la « *fonction témoin* ». Les valeurs de FRESA sont dans le même ordre de .48111, .19042 et .10028. Parmi l'ensemble des messages associés à l'université de Yale, les résumeurs automatiques (à l'exception de la « *fonction témoin* ») ont extrait le contenu suivant : « *'Report : Annie Le's Family Sues Yale University After Grad Student's Killing : NEW HAVEN, CT - The family of a slain Ya...'* » comme étant le tweet publié le plus pertinent au regard du sujet. Cette extraction s'explique facilement par le fait que l'étiquette du groupe de messages est pratiquement extraite telle quel du message. Si le système « *Artex* » a également pu extraire d'autres messages similaires relevant de cet événement, « *Cortex* » a de son coté retourné des messages relatifs à d'autres affaires sur le campus.

Afin de mieux comprendre les résultats obtenus par nos systèmes automatiques, intéressons-nous maintenant dans le tableau 6.4 aux couples (paquet de messages ; étiquette) que les systèmes ont été capables de manipuler. C'est à dire ceux pour lesquels les systèmes automatiques ont pu trouver au moins un message pertinent parmi les cinq messages qu'ils ont sélectionnés. Les faibles scores obtenus précédemment par « *Cortex* » et la « *fonction témoin* » sont ici illustrés par un nombre restreint de couples correctement appréhendés.

Ce niveau de performances assez bas est causé par les différentes caractéristiques des paquets de messages et des étiquettes de ceux-ci à partir desquelles sont formulées les requêtes. Nous avons par exemple remarqué sur l'ensemble des données que le système « *Cortex* » était bien plus efficace sur les paquets

Résumeurs		
CORTEX	ARTEX	FONCTION TÉMOIN
95	**950**	38

TABLE 6.4 – *Nombres de requêtes pour lesquelles les résumeurs obtiennent une MAP non-nulle.*

contenant plus de 20 messages ainsi que sur les requêtes plus vagues et difficiles à interpréter pour un expert. A l'inverse, «*Artex*» semble bien mieux fonctionner sur les petits groupes lorsqu'il s'agit de retrouver un petit nombre de messages correspondant à une requête directement extraite du contenu d'un des messages du groupe.

Enfin, il existe une limitation dans la manière dont la référence a été constituée. Un message ne peut appartenir qu'à un seul paquet. Or nous avons vu qu'il est possible que deux paquets pourtant différents dans la référence soient en fait similaires et traitent d'un même sujet. Un système automatique pourrait mélanger les messages de ces deux groupes et ainsi obtenir un très mauvais score alors qu'il répond pourtant correctement au besoin. Nous aurions pu à cet effet considérer une règle simple qui consisterait à itérativement écarter les messages les plus pertinents pour chaque requête afin d'éviter qu'ils soient pris en compte dans le calcul des pondérations pour les requêtes suivantes. Si cela permettrait d'augmenter les performances vis à vis de cette référence, cela nous détournerait de notre besoin initial de recherche d'information pertinente. De plus, une telle règle serait dépendante du point d'initialisation et de l'ordre dans lequel seraient traitées les requêtes.

6.6.2 Résumés de profils

Nous proposons de comparer les performances obtenues par les méthodes décrites dans le chapitre 5 sur l'ensemble des données avec les performances obtenues par ces mêmes méthodes sur un échantillon des données (méthodes notées «*Artex*» dans les tableaux 6.5 et 6.6). Cette fois encore, les résultats obtenus pour chaque domaine, pour la tâche de classement des utilisateurs sont ordonnés selon la MAP moyenne.

Il est intéressant d'observer que les systèmes travaillant à partir d'un échantillon réduit des messages publiés par un utilisateur obtiennent des résultats très similaires, parfois très légèrement supérieurs parfois très légèrement inférieurs, à ce qui était observable à partir de l'ensemble des données. L'absence de différences significatives entre les deux approches nous permet donc d'affirmer qu'il n'est pas nécessaire de considérer l'ensemble des publications d'un utilisa-

Approches			Automobile	Banques	Moyenne
UaD	Separated		**.803**	.626	**.714**
BoT	Separated	Sum	.779	.628	.703
BoT Artex	Separated	Sum	.774	**.633**	.703
UaD Artex	Separated		.782	.623	.702
BoT Artex	Separated	Count	.778	.612	.695
BoT	Separated	Count	.762	.592	.677
UaD	Joint		.735	.538	.636
UaD Artex	Joint		.722	.547	.634
BoT	Joint	Sum	.699	.526	.612
BoT	Joint	Count	.626	.504	.565

TABLE 6.5 – *Comparaisons des performances pour la tâche de classement d'utilisateurs par niveau d'influence, triées par MAP moyenne (les meilleurs scores sont indiqués en gras).*

Approches			Automobile	Banques	Moyenne
UaD	Separated		**.833**	**.751**	**.792**
UaD Artex	Separated		.829	.745	.787
BoT Artex	Separated	Sum	.820	.721	.770
BoT	Separated	Sum	.817	.709	.763
BoT Artex	Separated	Count	.796	.719	.757
BoT	Separated	Count	.786	.702	.744
UaD	Joint		.782	.682	.732
UaD Artex	Joint		.773	.672	.722
BoT	Joint	Count	.725	.641	.683
BoT	Joint	Sum	.725	.641	.683

TABLE 6.6 – *Performances pour la tâche de catégorisation d'utilisateur par niveau d'influence, triées par F-Score moyen (les meilleurs scores sont indiqués en gras).*

teur pour être en mesure de le profiler sur un aspect précis ou de le distinguer des autres utilisateurs. Ces résultats démontrent également que les méthodes choisies (notamment l'emploi du classique produit $TF_{i,d} \times IDF_i$) pour profiler les utilisateurs étaient également capables de ne retenir parmi les 600 messages que l'information nécessaire au traitement de la tâche.

6.6.3 Modélisation d'alerte

Nous considérons le cas où l'analyste définit un ensemble de thématiques D au travers duquel il souhaite suivre l'opinion et définir ses priorités pour analyser les tendances d'un flux de messages. Nous proposons un modèle combinant à la fois les thématiques définies par les spécialistes du « *Reputation Insti-*

tute » (voir tableau3.4) et les concepts « *d'alerte* » et « *d'importance* » dans le but de mieux suivre l'image de marque d'une entité. Le PLS-PM est utilisé pour trouver la pondération optimale (pour chaque paire de système-catégorie thématique) pour prédire une « *alerte* » conditionnelle à partir des probabilités données par l'ensemble des systèmes automatiques, basés sur l'analyse des contenus textuels, pour chacune des catégories thématiques et chaque niveau de priorité.

Étant donné une entité et un jeu de concepts à analyser, nous cherchons à modéliser l'impact du choix des thématiques à partir du flux de documents relatifs à l'entité. L'objectif est d'être à même de pouvoir expliquer à l'entité pourquoi cette thématique nécessite plus d'attention. Le modèle permet également de suivre dans le temps l'impact de chaque thématique sur l'image de marque de l'entité et par la combinaison de systèmes d'améliorer la détection des alertes en ayant la connaissance des thématiques à privilégier. Chaque thématique est modélisée comme une variable latente combinant plusieurs méthodes automatiques de catégorisation et entraînant l'un des deux concepts « *d'alerte* » et « *d'importance* », sous entendu que l'alerte induit l'importance.

A partir des données fournies dans le cadre de RepLab pour l'entraînement des systèmes nous pouvons faire apparaître une séparation des thématiques en deux groupes, celles qui vont directement induire « *l'alerte* » et celles qui au contraire semblent moins stratégiques (que l'on peut considérer comme non-importantes). Il faut toutefois garder à l'esprit qu'un tweet est généralement ambigu et peut correspondre à plusieurs thématiques, ces dernières comme l'innovation et le leadership sont vagues et trompeuses. La confiance accordée alors dans l'annotation ne peut être que faible.

Nous proposons également d'estimer le niveau de priorité « *alerte* » à partir de systèmes automatiques. « *L'importance* » est estimée à partir de la non-alerte (inverse de l'alerte, c'est à dire tout ce qui n'est pas alerte). Il apparaît également qu'il existe un modèle commun aux quatre domaines considérés dans le cadre du challenge. Cinq des thématiques induisent l'alerte, il s'agit de Gouvernance, Innovation, Leadership, Performances et Satisfaction. Le complément des deux restantes (Non-Citoyenneté et Non-Produits) implique l'importance (ou plus précisément la non-importance).

La figure 6.2 montre le modèle interne pour le domaine bancaire ainsi que les pondérations estimées par l'algorithme PLS-PM. L'ensemble des variables sont estimées en utilisant les scores de confiance des systèmes automatiques Cosine, Jaccard et SVM. Notons qu'une pondération négative n'implique pas obligatoirement une corrélation négative, l'algorithme signant les pondérations entre variables latentes de manière à maximiser la somme des corrélations internes entre les variables manifestes qui composent la variable latente.

Nous montrons ensuite sur la figure 6.3 le même modèle mais en intégrant

Domaine *Bancaire* avec référence

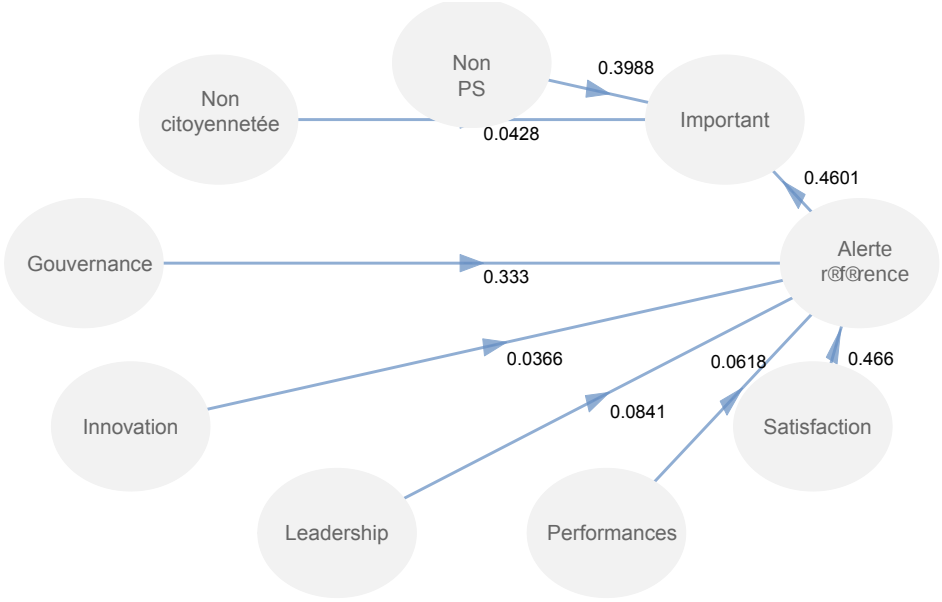


FIGURE 6.2 – *Modèle interne pour le domaine bancaire à partir de « l'alerte » de référence.*

cette fois-ci la référence de la variable « *Alerte* » issue des annotations et non son estimation fournie par les méthodes de catégorisation automatique. L'écart entre les corrélations obtenues pour ce modèle et le précédent est significatif (t-test p-value < 0.05). Toutefois, les thématiques restent classées dans le même ordre. La corrélation « *Pearson's product-moment* » entre l'estimation de « *l'alerte* » avec le modèle et la référence est .49 ce qui significativement haut (p-value $< 10^{-3}$) mais finalement pas meilleur que la seule estimation basée sur Cosine. Nous pouvons donc conclure concernant le domaine bancaire que la modélisation permet de hiérarchiser les thématiques selon le critère d'importance de l'analyste mais ne permet pas d'améliorer la qualité de l'estimation de cette priorité celle-ci étant déjà relativement bien estimée à partir d'un seul classifieur.

Comme le montre la figure 6.4, les résultats obtenus sont légèrement différents pour le domaine automobile. Le classement des thématiques n'est pas strictement identique car le modèle construit à partir de « *l'alerte* » estimée place la Satisfaction (coefficient à .466) devant Gouvernance (.333). Le modèle construit sur les références mesure l'impact de Innovation sur « *l'alerte* » à .03. Les deux classements restent cependant fortement corrélés (Kendall test :

Domaine *Bancaire* avec estimation

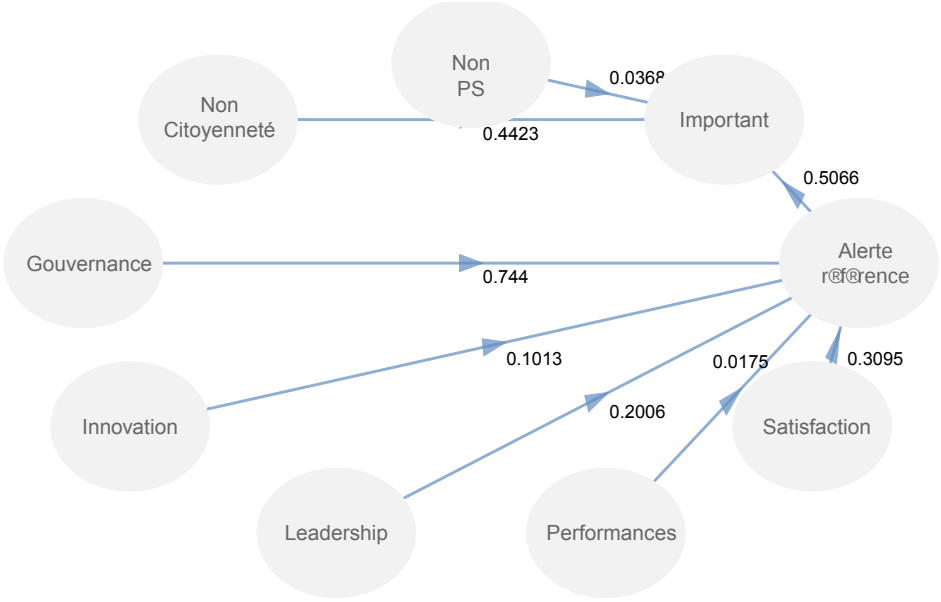


FIGURE 6.3 – *Modèle interne pour le domaine bancaire à partir de « l'alerte » estimée à partir des contenus textuels par les systèmes de catégorisation automatique.*

$80 < \tau < 1$, p-value< 0.05) et «*l'alerte*» estimée par le modèle reste fortement liée, à la référence (Interval de confiance à 95% entre .24 et .25 pour la corrélation de Pearson).

Ce lien est beaucoup plus faible sur les deux domaines restants (Artistes et Universités voir figure 6.5) (Pearson's product-moment corrélation à .15 et .11) principalement car les méthodes automatiques sont moins performantes sur ces domaines pour prédire correctement le niveau de priorité d'un message. Toutefois, «*l'alerte*» estimée par le modèle montre un renforcement de 10% de lien par rapport à la seule estimation de la meilleure méthode automatique. De plus, notons que pour ces deux domaines, aucune annotation thématique n'était disponible, les méthodes d'apprentissage automatique ont donc travaillé à partir des données disponibles pour les banques et les constructeurs automobile (les seuls pour lesquels les données étaient fournies mais également les plus concernés par la notion «*d'alerte*»). Les systèmes d'annotation automatique ont donc eu tendance à propager des règles associatives issues des banques (corrélations les plus fortes) aux autres domaines sans avoir toutefois inversé les relations preuve qu'un lien latent existe dans les données relatives à ces domaines.

Domaine *Automobile*

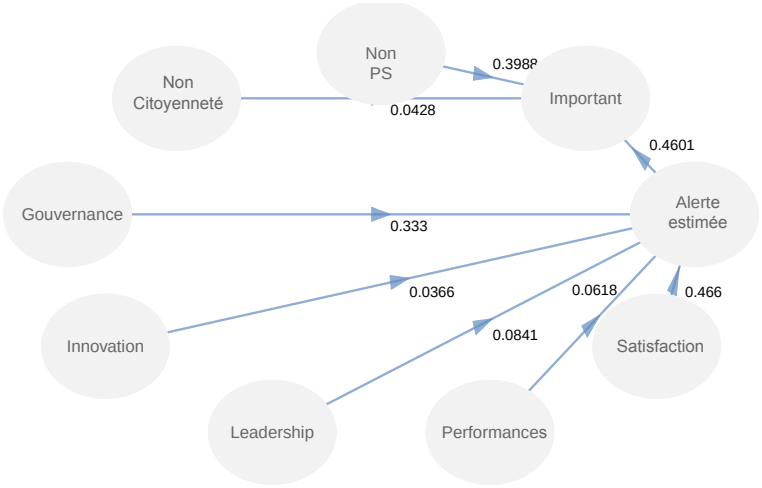


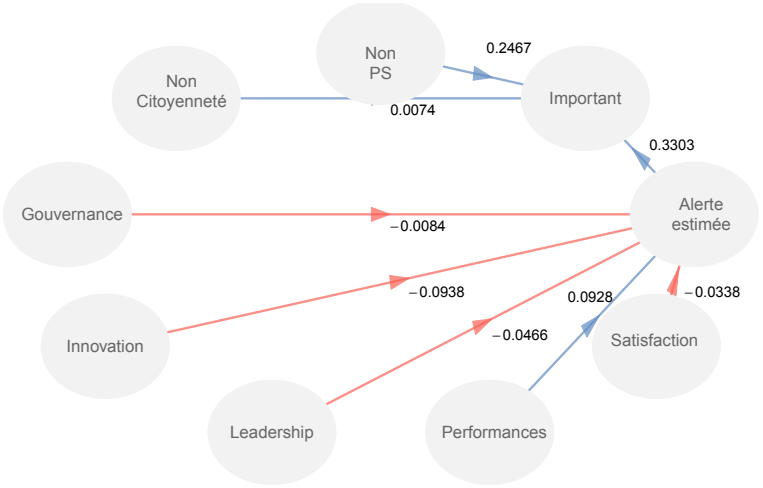
FIGURE 6.4 – *Modèle interne pour le domaine bancaire à partir de « l'alerte » estimée à partir des contenus textuels par les systèmes de catégorisation automatique.*

6.6.4 Modélisation d'influence

Nous avons vu dans le chapitre 5 que parmi les catégories de caractéristiques que nous avons retenues pour décrire un profil, le « *Style éditorial* » également utilisé (Ramírez-de-la Rosa et al., 2014) ne s'est pas montré pertinent pour la détection d'utilisateurs « *influants* » alors que la littérature insiste pourtant sur son pouvoir discriminant. Ce résultat nous amène à étudier l'existence de liens entre les différentes caractéristiques et le critère *d'influence* que ce dernier soit estimé par un système automatique ou déterminé par annotation manuelle.

Nous estimons que le cas d'étude de la détection d'influence est finalement fortement similaire à l'analyse de satisfaction clients telle que (Fornell, 1992) a pu la définir. Nous proposons un modèle conceptuel qui combine l'ensemble des catégories de caractéristiques que nous avons défini au début du chapitre 5. Notre objectif est d'arriver à expliquer en quoi l'exploitation des données issues de ces caractéristiques ont pu induire en erreur les méthodes d'apprentissage automatique et à l'inverse, chercher des relations fortes entre nos caractéristiques en tant que variables latentes et le critère d'influence que nous souhaitons modéliser. La modélisation que nous proposons se compose de quatre niveaux hiérarchiques : tout d'abord les caractéristiques (variables manifestes, observées et quantifiées), chacune composant une catégorie (variables latentes). Chaque catégorie est ensuite reliée soit à l'estimation de l'influence obtenue par

Domaine *Artistes*



Domaine *Université*

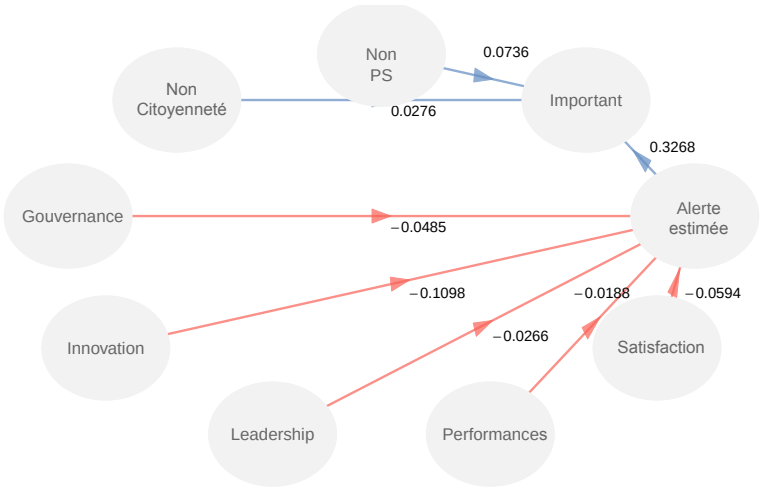


FIGURE 6.5 – *Modèle interne pour les domaines Artistes et Université à partir de « l'alerte » estimée à partir des contenus textuels par les systèmes de catégorisation automatique.*

le système automatique proposé dans le chapitre précédent soit à l'annotation d'influence de référence. Le système automatique se basant sur les contenus, nous ne cherchons pas intégrer sa corrélation aux catégories de caractéristiques

concernant le contenu textuel des messages. Ces catégories sont elles directement connectées à la référence étant donné qu'elles ont peu ou pas de lien avec les contenus. L'estimation du système est considérée comme un troisième niveau de modèle, ce dernier est également lié à la référence qu'il est censé estimer. Nous pouvons considérer que notre modèle se sépare en deux parties, les caractéristiques à partir desquelles il est possible d'induire l'influence telle qu'elle a été annotée et celles qui sont liées au système automatique [13] lui même capable de déduire l'influence d'un utilisateur.

La figure 6.6 montre par exemple les variables latentes représentant les catégories de caractéristiques *activité de publication* et *profil public* que nous avons décrites dans le chapitre précédent. La figure montre également les variables manifestes qui contribuent aux variables latentes.

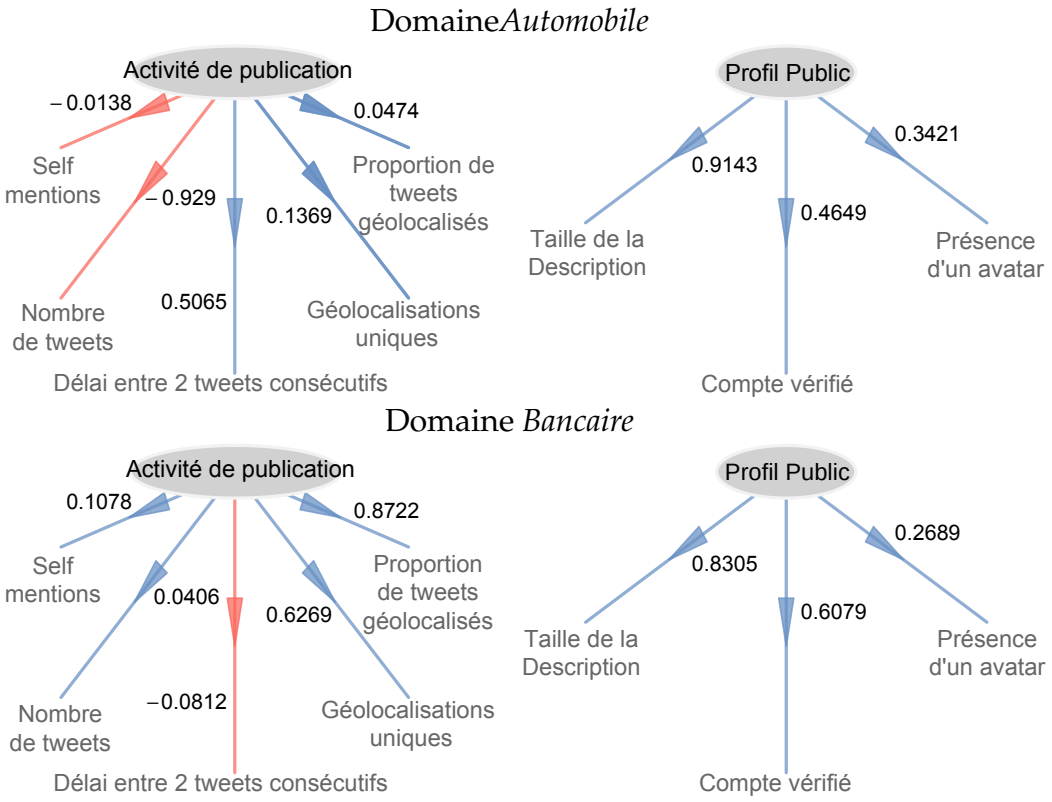


FIGURE 6.6 – *Pondérations interne des catégories liées à l'activité de publication (colonne de gauche) et Profil public (celle de droite) pour les modèles associés aux domaines Automobile (en haut) et bancaire (bas).*

La figure 6.6 montre que les corrélations associées aux différentes caracté-

13. Nous avons retenu la méthode ayant obtenu les meilleurs performances dans le chapitre précédent soit l'approche UaD basée sur traitement séparé des langues en considérant l'ensemble des publications d'un utilisateur

ristiques varient selon le domaine considéré. Pour le domaine Automobile par exemple, les contributions des caractéristiques (variables manifestes) liées aux «*self-mentions*» et à la géolocalisation des messages sont quasi nulles dans leur catégorie (variable latente). A l'inverse, les variables liées à un délai entre deux messages et au nombre total de publications atteignent des corrélations bien plus élevées (proche de 1 en valeur absolue pour le nombre de publications).

Pour le domaine relatif aux banques, les observations sont tout autres. Les aspects liés à la géolocalisation sont fortement corrélées à «*l'influence*» alors que les autres aspects montrent des pondérations très faibles. Les caractéristiques composant le profil public ont quant à elles un comportement stable. Quel que soit le domaine, la présence d'une image et l'indicateur d'authenticité du compte ne semblent pas impacter «*l'influence*» alors que la taille de la description semble plus importante. Ces valeurs indiquent un comportement type des utilisateurs influants qui sont plus susceptibles de mieux remplir leur profil afin de bien indiquer à leurs abonnés qui ils sont et ce qu'ils font.

Nous proposons maintenant de décrire le comportement des autres catégories et caractéristiques. Pour le domaine automobile les caractéristiques liées aux «*mots-dièses*» sont les principales composantes de la catégorie correspondant au style éditorial. Cette observation confirme l'intuition de (Aleahmad et al., 2014) sur la tendance des utilisateurs «*influants*» à rester à la pointe des dernières tendances et discussions du domaine. Pour le domaine des banques, les informations relatives aux liens [14] obtiennent les meilleures pondérations et semblent être les plus importantes de leur catégorie. Nous pensons que les futurs travaux sur le sujet devraient donc se focaliser sur cet aspect en allant justement chercher de l'information supplémentaire sur ces liens afin d'affiner la détection «*d'influence*». Ce contenu textuel complémentaire pourrait également être soumis à des systèmes automatiques de traitement de la langue afin d'en sélectionner uniquement l'information la plus importante. Parmi les aspects liés au champ lexical, la taille du vocabulaire obtient un score élevé, à l'inverse le nombre de «*Hapax*» (n'ayant été utilisé qu'une seule fois ou par un seul utilisateur (Ramírez-de-la Rosa et al., 2014)) ne semble important que pour le domaine automobile.

La figure 6.7 dépeint le modèle interne et ses pondérations c'est à dire les relations entre variables latentes et les différentes valeurs «*d'influence*» (estimation et référence mais également entre ces deux dernières). D'ailleurs «*l'influence*» estimée est fortement corrélée à la référence et ce pour les deux domaines bien que les valeurs observées soient plus proches de .5 que de 1. Cette observation confirme les bons résultats que nous avons obtenu dans le chapitre 5. Certaines catégories comme le profil public, les interactions avec le réseau

14. Nombre de liens uniques et nombre total de liens utilisés.

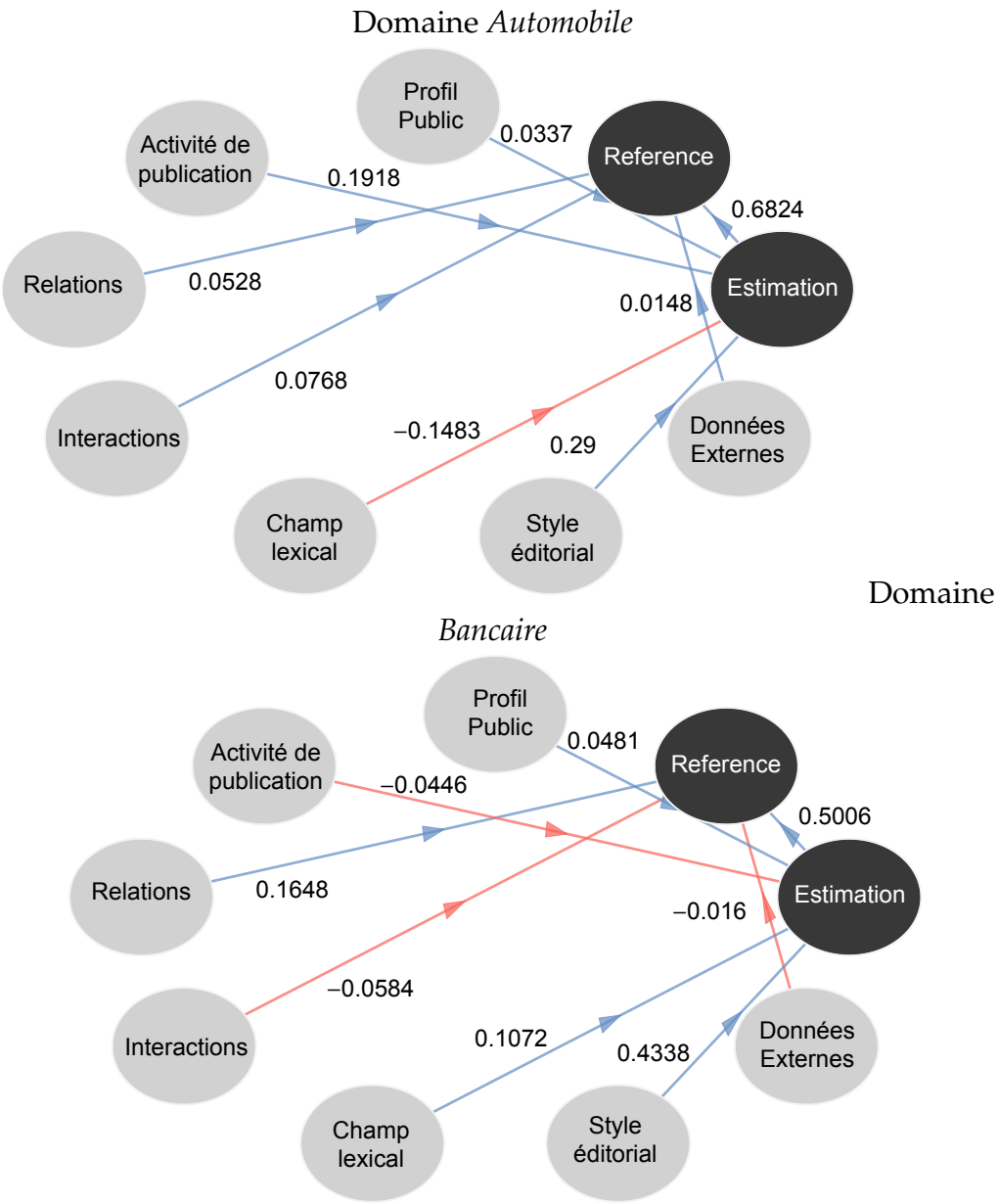


FIGURE 6.7 – *Corrélations entre les catégories de caractéristiques (variables latentes) et l'influence (réelle ou estimée) pour les domaines Automobile (haut) et bancaire (en bas).*

de relations et les ressources externes (score Klout) ont une corrélation quasi nulle quel que soit le domaine alors que nous avons pu voir que les corrélations internes aux catégories sont très élevées. Elles sont donc cohérentes entre elles mais non pertinentes vis a vis de notre analyse.

Quelques catégories comme l'activité de publication pour le domaine automobile, le réseau de relations pour les banques et celles liées aux champs

lexicaux dans les deux cas obtiennent des valeurs supérieures à .1 (en valeur absolue). Ces différences observées sur les pondérations nous permettent de confirmer la dépendance entre «*influence*» et domaine. «*L'influence*» se matérialise selon différents critères. Les aspects liés au style éditorial, au travers de leur catégorie de variables, obtiennent une corrélation bien plus forte et ce pour les deux domaines. Cette observation renforce l'intérêt d'utiliser des méthodes automatiques pour l'analyse des contenus plutôt que des méthodes dites «*était de l'art*» d'analyse de réseaux sociaux.

6.7 Conclusion

Nous nous avons vu dans ce chapitre la manière dont la littérature appréhende la question de la visualisation de l'information. S'il existe de nombreux outils de qualité tels que Gephi [15] permettant d'obtenir rapidement des représentations visuelles des données, l'adaptation aux caractéristiques du Web 2.0 (comme la mise à l'échelle) pose encore problème mais cela est très certainement lié à une mauvaise définition des axes d'analyse. Les outils de «*reporting*» et tableaux de bord bien plus populaires auprès des vielleurs s'affranchissent toutefois de cette question de définition d'axes d'analyse. Ces derniers se contentent le plus souvent d'offrir une vue consolidée mais restreinte de ce que contiennent les données. En effet, il n'est bien souvent question que de découpage de données et de projection selon différentes granularités temporelles des volumes de données correspondant à chaque étiquette d'opinion.

Nous avons montré dans le chapitre 5 qu'il était possible de caractériser les profils d'utilisateurs en fonction de leurs écrits. Dans ce chapitre, nous avons proposé de reprendre ces expérimentations en considérant cette fois-ci une méthode de sélection de contenus pour restreindre les données à notre disposition et nous placer dans un scénario où nous devrions fournir une aide à la décision alors que nous n'avons à notre disposition qu'un échantillon de la production d'un pseudo. Nous avons ensuite mis en perspective les différentes caractéristiques considérées par la littérature du domaine afin d'établir des relations entre les catégories de caractéristiques que nous avons définies. L'analyse des modèles que nous proposons permet à n'importe quel analyste, qu'il soit spécialiste ou non de prendre rapidement une décision sur l'intérêt pour son étude des caractéristiques qu'il souhaite analyser.

15. http://gephi.github.io/

En s'affranchissant des coûteuses étapes de conception de modèle ou de choix de modèles dans une banque de données spécialisée, nous proposons expérimentalement de vérifier directement la pertinence de nombreux modèles déduits directement des observations faites sur les données. Reste que la littérature est encore vague et limitée au sujet de l'évaluation de cette modélisation. (Jakobowicz, 2007) proposait de se limiter au cas où il était possible de définir une fonction à optimiser et d'utiliser cette dernière comme critère d'évaluation. Pour notre part, une fois qu'un modèle a émergé, nous proposons de rechercher notre fonction d'optimisation en déplaçant le problème dans un cadre probabiliste. L'étude théorique de cette fonction est toutefois encore en cours(Cossu et al., 2015c).

Chapitre 7

Conclusions et perspectives

Sommaire

Pour conclure ce manuscrit de thèse, nous résumons tout d'abord les travaux que nous avons présentés. Puis, nous terminons en synthétisant les principales perspectives de recherche ouvertes par ces travaux.

Est-il possible d'analyser de façon entièrement automatique l'image de marque d'entités sur le Web 2.0 selon les différentes attentes et le besoin d'information d'un utilisateur (attaché de communication, analyste de réputation) par ses définitions de pertinence et concepts à partir des contenus publiés sur les médias sociaux ? C'est la question principale qui a motivé le développement des méthodes que nous avons présentées dans cette thèse. Et c'est ce fil conducteur d'analyse des contenus que l'on aura retrouvé tout au long de cette thèse et l'utilisation dans nos approches de méthodes et algorithmes simples [1] et bien connues dans les domaines de l'apprentissage automatique. Nous nous sommes intéressés à la modélisation de cette image de marque selon les différents critères d'analyse proposés par les managers des différentes entités ou des spécialistes du domaine. Nous avons tout d'abord commencé par nous focaliser sur l'analyse des contenus textuels des messages publiés par les utilisateurs de médias sociaux. Nous avons décrit dans le chapitre 2 puis utilisé dans le chapitre 4 une approche statistique simple mais performante permettant à partir d'un échantillon réduit de connaissances d'experts (contenus annotés manuellement) (i) d'extraire des mots liés aux thématique, (ii) de les pondérer de façon à refléter l'importance de l'information qu'ils transmettent. C'est logiquement que nous avons ensuite analysé ces utilisateurs dans le chapitre 5 à partir justement de ce que ces derniers publiaient. La suite de cette conclusion reprend les résultats principaux que nous avons mis en valeur dans cette thèse.

7.1 Récapitulatif

Si nous pouvions intuitivement penser que l'utilisation de méthodes plus complexes recourant parfois même à des collections externes seraient beaucoup plus performantes pour traiter automatiquement les contenus, nous avons observé (et cette observation est partagée par la littérature (Amigó et al., 2014; Rangel et al., 2015)) que les différents systèmes présentaient globalement des performances homogènes. Chacun ayant ses particularités, présentant parfois de meilleures aptitudes pour certains types de contenus ou une catégorie spécifique. Ce plafond de performance parfois proche du taux d'accord que l'on pourrait observer si des annotateurs humains avaient traité les données relève de la qualité des sources de données que nous avons utilisées. Lorsque nous nous sommes intéressés aux «*a priori*» à la manière dont l'analyste (et

1. Ainsi que leur combinaison.

plus généralement l'homme) est conditionné et influençable durant l'annotation (contexte du message, contexte au moment de l'annotation). Nous avons observé que soumis à un même élément (contenu identique issu d'un même utilisateur) les analystes prenaient des décisions différentes. Cette observation nous rappelle la limite des conclusions que nous pouvons tirer de nos résultats. Nous avons également touché les limites de très larges collections de données, qui ont tendance à être difficilement appréhendables autrement qu'avec des méthodes automatiques. Ces méthodes permettent (parfois non sans difficulté) d'annoter automatiquement en amont d'immenses quantités des données avec une marge d'erreur. Mais qu'en serait-il alors du coût de l'analyse et la validation manuelle indispensable et complémentaire à cette annotation ?

Nous nous sommes intéressés aux éléments qui rentrent en ligne de compte dans la prise de décision de l'annotateur, de l'analyste. Comme observé dans la littérature par (Amigó et al., 2013a) l'émetteur du message est un indicateur puissant du sens de ce message. (Peetz, 2015) indique que l'autorité de l'utilisateur dans son domaine est un indice fort pour déterminer la tonalité et l'importance d'un message. Nous avons proposé avec succès dans le chapitre 5 des méthodes permettant d'établir de manière automatique quel était le niveau d'autorité « *hors-ligne* » d'un utilisateur à partir de données qu'il publiait « *en-ligne* ». Ces mêmes méthodes nous permettaient également de séparer les utilisateurs influents de ceux qui ne le sont pas. Quand tout cela n'est pas suffisant, l'annotateur se retourne vers des moteurs de recherche d'information afin de remettre le contenu en contexte et ainsi mieux définir la tonalité ou la thématique du message. Nous avons montré dans le chapitre 4 que dans ce dernier cas précisément, l'information additionnelle permet de mieux séparer les messages relevant de thématiques bien définies de ceux relevant de thématiques encore floues ou émergentes.

La plupart de expérimentations proposées dans cette thèse étaient motivées par les besoins du projet **Imagiweb** mais également par notre volonté d'établir une relation de confiance entre les chercheurs en TAL chargés d'implémenter les algorithmes pour annoter automatiquement les données et les chercheurs en sciences politiques, spécialistes de leur domaine et annotateurs vérifiant la qualité des annotations. L'objectif était de proposer des méthodes permettant de s'affranchir de cette coûteuse vérification en intégrant directement dans le fonctionnement des algorithmes cette connaissance inestimable des spécialistes. Même s'il faut (certes peu) des exemples d'annotations manuels pour que les systèmes fonctionnent, de grands pas ont été fait. Nous avons vu au cours de cette thèse qu'il était possible d'automatiser des éléments qui constituent ce savoir faire du spécialiste.

La fin de cette conclusion offre des perspectives d'évolution et de travaux futurs.

7.2 Perspectives

Comme nous venons de le voir, cette thèse nous aura permis de tirer plusieurs leçons concernant l'analyse de l'image de marques d'entités sur le Web 2.0. Nous avons bien sûr rencontré plusieurs problèmes tout au long de cette thèse et certains sont restés non résolus.

Les flux d'activité sur Twitter sont très dynamiques, ce qui est dit à un instant t au sujet d'une entité a de forte chance d'évoluer dans le temps. Ce qui était important aujourd'hui le sera t'il encore demain ? Celui qui était considéré comme la référence sur la question sera-t-il toujours le seul à l'être ? Comme on le sait, l'actualité évolue rapidement d'autant plus sur les réseaux sociaux et le vocabulaire évolue en fonction de cette actualité [2]. Cette modification des perceptions des internautes au fil du temps est un phénomène bien connu. Il est alors difficile de s'assurer que les modèles que nous venons d'apprendre à partir des observations disponibles jusqu'à présent seront encore valables pour les événements à venir à plus long terme. Un processus «*d'adaptation*» s'avère donc indispensable.

La littérature de l'apprentissage actif (Settles, 2012) indique qu'après validation il est possible d'intégrer de nouveaux éléments dans la base de connaissances qui permet de créer les modèles. A la différence de (Peetz, 2015), nous nous sommes intéressés à cette possibilité dans nos expériences et en proposant de nous affranchir de cette coûteuse étape de validation en utilisant directement une sélection d'hypothèses des systèmes. Il s'agissait d'un des objectifs qui occupait une place de choix, que nous avons à plusieurs reprises voulu atteindre sans jamais y parvenir. Afin de tenir compte de la nouveauté et des tendances d'opinions, chaque fois que nous avons essayé d'adapter automatiquement les paramètres des modèles numériques que nous utilisons pour effectuer des tâches de catégorisation thématique ou de détection de polarité, nous avons observé à regret une dégradation des résultats au lieu d'une amélioration. Ceci est d'autant plus regrettable que ces modèles s'appuient sur les termes contenus dans les messages traités et de plus, un des avantages de la méthode que nous proposions est qu'elle est entièrement automatique, et ne requiert donc aucune phase d'entraînement préalable ni supervision. Nous aurions également pu envisager de modifier la pondération des mots en fonction de la décision, de la confiance du classifieur mais également de l'intérêt porté à la thématique ou au sujet du message. Comme nous avons à notre disposition un grand nombre d'hypothèses générées automatiquement par différents sys-

2. Événements majeurs (la catastrophe de Fukushima a entraîné l'association du nom de cette ville avec l'entreprise EDF), émergences de nouvelles affaires concernant les hommes politiques (Khadafi, Gayet etc.), reprises de slogan («*le mariage pour tous*» est devenu «*la manif pour tous*»).

tèmes en complément de ce que nous avons proposé dans (Cossu et al., 2015b) nous pourrions apprendre un classifieur supervisé pouvant décider de l'hypothèse à retenir. La validation des modèles et le choix des hypothèses à retenir pourrait également passer par la définition d'une fonction d'optimisation assurant la cohérence d'un modèle au regard des données à partir desquelles il est généré. Toutefois, l'étude théorique de cette fonction est encore en cours (Cossu et al., 2015c).

Nous envisageons également d'étudier les comportements des annotateurs, comment pour définir l'image de marque ces derniers interagissent avec les experts à la fois du domaine (politologues) et des réseaux sociaux. Nous avons vu dans le chapitre 4 que l'annotateur, fût-il expert, peut par facilité ou inattention se laisser influencer par des hypothèses qui seraient « *acceptables* »[3] au regard d'un contenu. Ceci est d'ailleurs totalement paradoxal, alors que les analystes n'accordent aucune confiance aux méthodes automatiques ils attendent de ces mêmes méthodes qu'elles soient les plus fiables possibles pour les assister. Heureusement, les méthodes que nous avons proposées, mêmes imparfaites, rendent la compréhension des mécanismes sociaux liés à l'image de marque moins fastidieuse. Nous sommes toutefois dépendant des intentions de ces experts et de ce qu'ils souhaitent faire émerger des données. A ce titre, il n'existe aucune étude (la validation manuelle « *objective* » resterait à définir) portant spécifiquement sur les différences entre les intentions de l'analyse et ce qui émerge réellement des données. Les méthodes automatiques d'analyse de contenus touchent à tous les aspects. Les algorithmes retrouvent, organisent et diffusent l'information, dans notre société conditionnée par l'abondance d'information. Il n'y a qu'un pas vers la configuration d'un système pour filtrer et contrôler l'information et discréditer l'autorité « *en-ligne* » des sources qui seraient gênantes « *hors-ligne* ».

Notre travail est resté éloigné de certaines attentes des analystes, nos algorithmes ne doivent pas seulement offrir des prédictions fiables mais également proposer des interactions avec l'utilisateur. Il est primordial d'envisager un dialogue avec l'analyste. Cela doit devenir la pré-occupation majeur des chercheurs sur le sujet pour permettre aux systèmes de s'adapter automatiquement d'autant plus que les analystes seraient prêt à passer du temps pour développer un assistant automatique des plus performant. « *est-ce que ce volume de messages autour de ce sujet est important ou s'agit-il d'un non événement ?* » intégrer ce genre d'information dans le fonctionnement d'un système d'analyse d'image de marque et d'annotation automatique permettrait peut-être même de dépasser bien sûr toutes les approches existantes mais aussi les attentes de l'expert et lui montrer une autre exécution de son processus métier. Toutefois

3. C'est à dire ne convenant pas tout à fait mais n'étant parallèlement pas véritablement erronées.

ce niveau d'interactions requiert des capacités d'intelligence artificielle que les spécialistes de l'apprentissage automatique actif sont encore loin de maîtriser. Parmi les attentes de l'analyste que nous sommes toujours loin de satisfaire, nous citons la définition des priorités (Mather et Sutherland, 2011) qui impliquerait que des futurs travaux élargissent le spectre d'application en collaborant par exemple avec d'autres chercheurs en sciences cognitives et humanités numériques pour mieux formaliser les nouveaux besoins qu'engendre ce monde du numérique.

Enfin, d'une manière plus générale, s'il est un bien un facteur plus limitant c'est ce que pensent les utilisateurs de médias sociaux que nous analysons quant au respect de leur vie privée et de leur liberté. Quand bien même leurs écrits sont disponibles publiquement, les législations évoluent rapidement sur la question. Ce qui est toléré aujourd'hui d'une manière massive comme des statistiques visant à connaître l'opinion d'une catégorie de la population, ces dernières sont obtenues à partir de données individuelles et si les pouvoirs publics s'autorisent à analyser précisément ces données [4] le cadre est encore relativement flou autour des pratiques de sociétés privées spécialisées dans le domaine.

4. http://goo.gl/erGuuZ

Bibliographie

(Abascal-Mena et al., 2015) R. Abascal-Mena, R. Lema, & F. Sèdes, 2015. Detecting sociosemantic communities by applying social network analysis in tweets. *Social Network Analysis and Mining* 5(1), 1–17. 88

(Abboute et al., 2014) A. Abboute, Y. Boudjeriou, G. Entringer, J. Azé, S. Bringay, & P. Poncelet, 2014. Mining twitter for suicide prevention. Dans E. Métais, M. Roche, & M. Teisseire (Eds.), *Natural Language Processing and Information Systems - 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings*, Volume 8455 de *Lecture Notes in Computer Science*, 250–253. Springer. 79

(Abracos et Lopes, 1997) J. Abracos & G. P. Lopes, 1997. Statistical Methods for Retrieving Most Significant Paragraphs in Newspaper Articles. Dans I. Mani & M. T. Maybury (Eds.), *ACL/EACL97-WS*, Madrid, Spain. 107

(Aggarwal, 2011) C. Aggarwal, 2011. Social network data analytics. 4, 5

(Al Zamal et al., 2012) F. Al Zamal, W. Liu, & D. Ruths, 2012. Homophily and latent attribute inference : Inferring latent attributes of Twitter users from neighbors. Dans les actes de *ICWSM*. 80, 86, 89

(Aleahmad et al., 2014) A. Aleahmad, P. Karisani, M. Rahgozar, & F. Oroumchian, 2014. University of tehran at replab 2014. Dans les actes de *4th International Conference of the CLEF initiative*. 88, 95, 97, 126

(Amigó et al., 2014) E. Amigó, J. Carrillo-de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, E. Meij, M. de Rijke, & D. Spina, 2014. Overview of replab 2014 : author profiling and reputation dimensions for online reputation management. Dans les actes de *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, 307–322. 40, 41, 44, 50, 74, 95, 100, 105, 131

(Amigó et al., 2013a) E. Amigó, J. C. De Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martín, E. Meij, M. De Rijke, & D. Spina, 2013a. Overview of replab 2013 : Evaluating online reputation monitoring systems. Dans les actes de *CLEF 2013*. 17, 33, 40, 45, 70, 105, 132

(Amigó et al., 2013b) E. Amigó, J. Gonzalo, & F. Verdejo, 2013b. A general evaluation measure for document organization tasks. Dans les actes de *Proc. of the 36th international SIGIR conference on Research and development in information retrieval*. 33

(Anagnostopoulos et al., 2008) A. Anagnostopoulos, R. Kumar, & M. Mahdian, 2008. Influence and correlation in social networks. Dans les actes de *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 7–15. ACM. 82

(Anger et Kittl, 2011) I. Anger & C. Kittl, 2011. Measuring influence on Twitter. Dans les actes de *i-KNOW*, 1–4. 80, 87, 92

(Bakshy et al., 2012) E. Bakshy, I. Rosenn, C. Marlow, & L. Adamic, 2012. The role of social networks in information diffusion. Dans les actes de *Proceedings of the 21st international conference on World Wide Web*, 519–528. ACM. 5

(Bellot et al., 2014) P. Bellot, V. Moriceau, J. Mothe, E. SanJuan, & X. Tannier, 2014. Overview of INEX tweet contextualization 2014 track. Dans L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, Volume 1180 de *CEUR Workshop Proceedings*, 494–500. CEUR-WS.org. 46, 68, 70

(Benevenuto et al., 2010) F. Benevenuto, F. Magno, T. Rodrigues, & V. Almeida, 2010. Detecting spammers on Twitter. Dans les actes de *CEAS*. 79, 85, 86, 87, 89, 90, 92

(Bengio et al., 2003) Y. Bengio, R. Ducharme, & P. Vincent, 2003. A neural probabilistic language model. *Journal of Machine Learning Research 3*, 1137–1155. 47

(Blum et Mitchell, 1998) A. Blum & T. Mitchell, 1998. Combining labeled and unlabeled data with co-training. Dans les actes de *Proceedings of the eleventh annual conference on Computational learning theory*, 92–100. ACM. 19

(Bollen et al., 2011) J. Bollen, H. Mao, & X. Zeng, 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2(1), 1–8. 16

(Bonnefoy et al., 2013) L. Bonnefoy, V. Bouvier, & P. Bellot, 2013. A weakly-supervised detection of entity central documents in a stream. Dans les actes de *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 769–772. ACM. xviii, 21, 22

(Boudin, 2008) F. Boudin, 2008. *Exploration d'approches statistiques pour le résumé automatique de texte*. doctorat en informatique, Laboratoire Informatique d'Avignon, Université d'Avignon et des Pays de Vaucluse, Avignon, France. 107

(Boyadjian, 2014) J. Boyadjian, 2014. Twitter, un nouveau «baromètre de l'opinion publique» ? *Participations* (1), 55–74. 4

(Boyd et al., 2010) D. Boyd, S. Golder, & G. Lotan, 2010. Tweet, tweet, retweet : Conversational aspects of retweeting on twitter. Dans les actes de *HICSS*, 1–10. 87

(Brun et Roux, 2014) C. Brun & C. Roux, 2014. Décomposition des «hash tags» pour l'amélioration de la classification en polarité des «tweets». *TALN, Juillet 146*. 7

(Carrillo-de Albornoz et al., 2014) J. Carrillo-de Albornoz, E. Amigó, D. Spina, & J. Gonzalo, 2014. Orma : A semi-automatic tool for online reputation monitoring in twitter. Dans les actes de *Advances in Information Retrieval*, 742–745. Springer. 36, 41

(Chen, 2006) C. Chen, 2006. *Information visualization : Beyond the horizon*. Springer Science & Business Media. 105

(Chen et Goodman, 1996) S. F. Chen & J. Goodman, 1996. An empirical study of smoothing techniques for language modeling. Dans les actes de *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 310–318. Association for Computational Linguistics. 26

(Cheng et Lee, 2010) J. Cheng, Z. Caverlee & K. Lee, 2010. You are where you tweet : a content-based approach to geo-locating twitter users. Dans les actes de *CIKM*, 759–768. 80

(Chu et al., 2012) Z. Chu, S. Gianvecchio, H. Wang, & S. Jajodia, 2012. Detecting automation of Twitter accounts : Are you a human, bot, or cyborg ? *IEEE Transactions on Dependable and Secure Computing* *9*(6), 811–824. 80, 84, 85, 89, 90, 91, 92

(Conover et al., 2011) M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, & F. Menczer, 2011. Predicting the political alignment of Twitter users. Dans les actes de *IEEE SocialCom*, 192–199. 80, 87, 88, 89

(Cossu et al., 2013) J. Cossu, B. Bigot, L. Bonnefoy, M. Morchid, X. Bost, G. Senay, R. Dufour, V. Bouvier, J. Torres-Moreno, & M. El-Bèze, 2013. Lia@replab 2013. Dans les actes de *CLEF*. 53, 57, 58, 60, 76, 154, 155

(Cossu et al., 2014) J. Cossu, B. Bigot, L. Bonnefoy, & G. Senay, 2014. Towards the improvement of topic priority assignment using various topic detection methods for e-reputation monitoring on twitter. Dans les actes de *Natural Language Processing and Information Systems - 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings*, 154–159. 12

(Cossu et al., 2015a) J.-V. Cossu, L. Bonnefoy, X. Bost, & M. El-Bèze, 2015a. How to merge three different methods for information filtering ? Rapport technique. 56

(Cossu et al., 2015b) J.-V. Cossu, N. Dugué, & V. Labatut, 2015b. Detecting real-world influence through Twitter. Dans les actes de *ENIC*. 12, 80, 81, 83, 88, 96, 97, 98

(Cossu et al., 2014) J.-V. Cossu, M. El-Bèze, J.-M. Torres Moreno, & E. Sanjuan, 2014. E-reputation monitoring on twitter with active learning automatic annotation. Rapport technique. 20

(Cossu et al., 2013) J.-V. Cossu, J. Gaillard, J.-M. Torres-Moreno, & M. El-Bèze, 2013. Contextualisation de messages courts : lâimportance des métadonnées. 157, 162

(Cossu et al., 2014) J.-V. Cossu, K. Janod, E. Ferreira, J. Gaillard, & M. El-Bèze, 2014. Lia@ replab 2014 : 10 methods for 3 tasks. Dans les actes de *4th International Conference of the CLEF initiative*, Sheffield (UK). 49, 93, 94, 95, 96

(Cossu et al., 2015a) J.-V. Cossu, K. Janod, E. Ferreira, J. Gaillard, & M. El-Bèze, 2015a. Nlp-based classifiers to generalize experts assessments in e-reputation. Dans les actes de *Experimental IR meets Multilinguality, Multimodality, and Interaction*. 12, 49, 93, 94, 96

(Cossu et al., 2015b) J.-V. Cossu, E. San-Juan, J.-M. Torres-Moreno, & M. El-Bèze, 2015b. Automatic classification and pls-pm modeling for profiling reputation of corporate entities on twitter. Dans les actes de *Natural Language Processing and Information Systems - 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015. Proceedings*. 13, 134

(Cossu et al., 2015c) J.-V. Cossu, E. San-Juan, J.-M. Torres-Moreno, & M. El-Bèze, 2015c. Multi-dimensional reputation modeling using micro blog contents. Dans les actes de *Foundations of Intelligent Systems - The 22th International Symposium on Methodologies for Intelligent Systems, ISMIS 2015, October 21-23, 2015, Lyon, France*. Springer. 129, 134

(Cossu et al., 2013) J.-V. Cossu, J.-M. Torres-Moreno, & M. El-Bèze, 2013. Recherche et utilisation d'entités nommées conceptuelles dans une tâche de catégorisation. Dans les actes de *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles*, Les Sables d'Olonne, France, 715–722. Association pour le Traitement Automatique des Langues. 21, 157, 165

(Cossu et al., 2015) J.-V. Cossu, J.-M. Torres-Moreno, E. San-Juan, & M. El-Bèze, 2015. Intweetive text summarization. Dans les actes de *Mexican International Conference on Artificial Intelligence (MICAI)*. 13

(Crammer et Singer, 2002) K. Crammer & Y. Singer, 2002. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research 2*, 265–292. 25

(Croft et Harper, 1979) W. B. Croft & D. J. Harper, 1979. Using probabilistic models of document retrieval without relevance information. *Journal of documentation 35*(4), 285–295. 20

(da Cunha et al., 2007) I. da Cunha, S. Fernández, P. Velázquez-Morales, J. Vivaldi, E. SanJuan, & J. Torres-Moreno, 2007. A new hybrid summarizer based on vector space model, statistical physics and linguistics. Dans les actes de *MICAI*, 872–882. 107

(Damak et al., 2013) F. Damak, K. Pinel-Sauvagnat, M. Boughanem, & G. Cabanac, 2013. Effectiveness of state-of-the-art features for microblog search. Dans les actes de *The 28th ACM Symposium on Applied Computing*. 21

(Das et Martins, 2007) D. Das & A. F. Martins, 2007. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU 4*, 192–195. 107

(de Choudhury et al., 2012) M. de Choudhury, N. Diakopoulos, & M. Naaman, 2012. Unfolding the event landscape on Twitter : classification and exploration of user categories. Dans les actes de *ACM CSCW*, 241–244. 80, 87, 88, 89

(de Silva et Riloff, 2014) L. de Silva & E. Riloff, 2014. User type classification of tweets with implications for event recognition. Dans les actes de *Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 98–108. 80, 88, 89

(Derczynski et al., 2015) L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, & K. Bontcheva, 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*. 8, 20

(Dugué et Perez, 2014) N. Dugué & A. Perez, 2014. Social capitalists on Twitter : detection, evolution and behavioral analysis. *Social Network Analysis and Mining 4*(1), 1–15. Springer. 80, 82, 86, 90

(Dugué et al., 2015) N. Dugué, A. Perez, M. Danisch, F. Bridoux, A. Daviau, T. Kolubako, S. Munier, & H. Durbano, 2015. A reliable and evolutive web

application to detect social capitalists. Dans les actes de *IEEE/ACM ASONAM Exhibits and Demos*. 80, 91

(Ellison et al., 2007) N. B. Ellison et al., 2007. Social network sites : Definition, history, and scholarship. *Journal of Computer-Mediated Communication* *13*(1), 210–230. 7

(Ermakova, 2014) L. Ermakova, 2014. Irit at inex 2014 : Tweet contextualization track. Dans les actes de *4th International Conference of the CLEF initiative*, Sheffield (UK). 69

(Fornell, 1992) C. Fornell, 1992. A national customer satisfaction barometer : the swedish experience. *Journal of Marketing*, 6–21. 115, 123

(Frank et al., 2012) J. Frank, M. Kleiman-Weiner, D. Roberts, F. Niu, C. Zhang, & C. Ré, 2012. Building an entity-centric stream filtering test collection for trec 2012. *Proceedings of The 21th TREC*. 21

(Gaillard, 2014) J. Gaillard, 2014. *Recommender Systems : Dynamic Adaptation and Argumentation*. doctorat en informatique, Laboratoire Informatique d'Avignon, Université d'Avignon et des Pays de Vaucluse, Avignon, France. 17

(Gârbacea et al., 2014) C. Gârbacea, M. Tsagkias, & M. de Rijke, 2014. Detecting the reputation polarity of microblog posts. ECAI. 53

(Gerlitz et Rieder, 2013) C. Gerlitz & B. Rieder, 2013. Mining one percent of twitter : Collections, baselines, sampling. *M/C Journal* *16*(2). 4

(Ghosh et al., 2012) S. Ghosh, B. Viswanath, F. Kooti, N. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, & K. Gummadi, 2012. Understanding and combating link farming in the Twitter social network. Dans les actes de *WWW*, 61–70. 79, 80, 90

(Gourion et Josselin, 2012) D. Gourion & D. Josselin, 2012. Aide à la décision robuste pour la localisation d'un centre de traitement des déchets. comparaison de méthodes d'analyse multicritères. Dans les actes de *Annales de l'ISUP*, Volume 56, 17–36. Institut de statistique de l'Université de Paris. 28

(Hangya et Farkas,) V. Hangya & R. Farkas. Filtering and polarity detection for reputation management on tweets. Dans les actes de *CLEF 2013*. 56, 57, 155

(Hazen et al., 2007) T. J. Hazen, F. Richardson, & A. Margolis, 2007. Topic identification from audio recordings using word and phone recognition lattices. Dans les actes de *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, 659–664. IEEE. 60

(Henseler, 2010) J. Henseler, 2010. On the convergence of the partial least squares path modeling algorithm. *Computational Statistics* *25*(1), 107–120. 112

(Hovy et Lin, 1999) E. Hovy & C. Y. Lin, 1999. Automated Text Summarization in SUMMARIST. Dans I. Mani & M. T. Maybury (Eds.), *Advances in Automatic Text Summarization*, 81–94. The MIT Press. 107

(Huang et al., 2014) W. Huang, I. Weber, & S. Vieweg, 2014. Inferring nationalities of Twitter users and studying inter-national linking. Dans les actes de *ACM Hypertext*. 80, 84, 85, 91

(Jakobowicz, 2007) E. Jakobowicz, 2007. *Contributions aux modèles d'équations structurelles à variables latentes*. Thèse de Doctorat, Conservatoire national des arts et metiers-CNAM. 111, 112, 115, 129

(Jansen et al., 2009) B. J. Jansen, M. Zhang, K. Sobel, & A. Chowdury, 2009. Twitter power : Tweets as electronic word of mouth. *Journal of the American society for information science and technology* *60*(11), 2169–2188. 15

(Java et al., 2007) A. Java, X. Song, T. Finin, & B. Tseng, 2007. Why we twitter : understanding microblogging usage and communities. Dans les actes de *WebKDD/SNA-KDD*, 56–65. 85, 86, 89

(Ježek et Steinberger, 2008) K. Ježek & J. Steinberger, 2008. Automatic text summarization (the state of the art 2007 and new challenges). *Proceedings of Znalosti 2008*, 1–12. 107

(Joachims, 1998) T. Joachims, 1998. *Text categorization with support vector machines : Learning with many relevant features*. Springer. 25

(Jungherr et al., 2012) A. Jungherr, P. Jürgens, & H. Schoen, 2012. Why the pirate party won the german election of 2009 or the trouble with predictions : A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im 'predicting elections with twitter : What 140 characters reveal about political sentiment'. *Social Science Computer Review* *30*(2), 229–234. 16

(Khalid et al., 2008) M. A. Khalid, V. Jijkoun, & M. De Rijke, 2008. The impact of named entity normalization on information retrieval for question answering. Dans les actes de *Advances in Information Retrieval*, 705–710. Springer. 20

(Kim et al., 2014) H. Kim, K. Beznosov, & E. Yoneki, 2014. Finding influential neighbors to maximize information diffusion in twitter. Dans les actes de *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, 701–706. International World Wide Web Conferences Steering Committee. 82

(Kim et al., 2015) Y.-M. Kim, J. Velcin, S. Bonnevay, & M.-A. Rizoiu, 2015. Temporal multinomial mixture for instance-oriented evolutionary clustering. Dans les actes de *Advances in Information Retrieval*. 93

(Klout, 2015) Klout, 2015. Klout, the standard for influence. 90

(Kontostathis et al., 2010) A. Kontostathis, L. Edwards, & A. Leatherman, 2010. Text mining and cybercrime. *Text Mining : Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK*. 8

(Koolen et Kamps, 2010) M. Koolen & J. Kamps, 2010. The importance of anchor text for ad hoc search revisited. Dans les actes de *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 122–129. ACM. 20

(Kred, 2015) Kred, 2015. Kred story. 90

(Kupiec et al., 1995) J. Kupiec, J. O. Pedersen, & F. Chen, 1995. A Trainable Document Summarizer. Dans les actes de *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 68–73. 107

(Laasby, 2014) G. Laasby, 2014. Blocking fake Twitter followers and spam accounts just got easier. 89

(Lamontagne et Abi-Zeid, 2006) L. Lamontagne & I. Abi-Zeid, 2006. Combining multiple similarity metrics using a multicriteria approach. Dans les actes de *Advances in Case-Based Reasoning*, 415–428. Springer. 28

(Laniado et Mika, 2010) D. Laniado & P. Mika, 2010. Making sense of twitter. Dans les actes de *The Semantic Web–ISWC 2010*, 470–485. Springer. 20

(Lee et al., 2010) K. Lee, J. Caverlee, & S. Webb, 2010. Uncovering social spammers : social honeypots + machine learning. Dans les actes de *ACM SIGIR*, 435–442. 79, 90

(Lee et al., 2011) K. Lee, B. D. Eoff, & J. Caverlee, 2011. Seven months with the devils : A long-term study of content polluters on Twitter. Dans les actes de *ICWSM*. 79, 85, 90, 91, 92

(Lee et al., 2013) K. Lee, P. Tamilarasan, & J. Caverlee, 2013. Crowdturfers, campaigns, and social media : Tracking and revealing crowdsourced manipulation of social media. Dans les actes de *ICWSM*. 80, 86, 88, 90

(Li et al., 2011) F. Li, N. Liu, H. Jin, K. Zhao, Q. Yang, & X. Zhu, 2011. Incorporating reviewer and product information for review rating prediction. Dans les actes de *IJCAI*, Volume 11, 1820–1825. 7

(Lin et al., 2006) C.-Y. Lin, G. Cao, J. Gao, & J.-Y. Nie, 2006. An information-theoretic approach to automatic evaluation of summaries. Dans les actes de *Conference on Human Language Technology Conference of the North American Chapter*, Morristown, NJ, USA, 463–470. ACL. 114

(Livne et al., 2011) A. Livne, M. P. Simmons, E. Adar, & L. A. Adamic, 2011. The party is over here : Structure and content in the 2010 election. *ICWSM 11*, 17–21. 16

(Lomena et Ostenero, 2014) J. J. M. Lomena & F. L. Ostenero, 2014. Uned at clef replab 2014 : Author profiling. Dans les actes de *4th International Conference of the CLEF initiative*. 95, 96

(Losada et Azzopardi, 2008) D. E. Losada & L. Azzopardi, 2008. An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval 11*(2), 109–138. 26

(Louis et Nenkova, 2009) A. Louis & A. Nenkova, 2009. Automatically Evaluating Content Selection in Summarization without Human Models. Dans les actes de *Empirical Methods in Natural Language Processing*, Singapore, 306–314. 114

(Mackie et al., 2014a) S. Mackie, R. McCreadie, C. Macdonald, & I. Ounis, 2014a. Comparing algorithms for microblog summarisation. Dans les actes de *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, 153–159. Springer. 107, 114

(Mackie et al., 2014b) S. Mackie, R. McCreadie, C. Macdonald, & I. Ounis, 2014b. On choosing an effective automatic evaluation metric for microblog summarisation. Dans les actes de *Proceedings of the 5th Information Interaction in Context Symposium*, 115–124. ACM. 107, 114

(Mahmud et al., 2012) J. Mahmud, J. Nichols, & C. Drews, 2012. Where is this tweet from ? inferring home locations of Twitter users. Dans les actes de *ICWSM*. 80

(Makazhanov et Rafiei, 2013) A. Makazhanov & D. Rafiei, 2013. Predicting political preference of Twitter users. Dans les actes de *IEEE/ACM ASONAM*, 298–305. 80, 89

(Makkonen et al., 2004) J. Makkonen, H. Ahonen-Myka, & M. Salmenkivi, 2004. Simple semantics in topic detection and tracking. *Information retrieval 7*(3-4), 347–368. 8, 11

(Manning et Schütze, 1999) C. D. Manning & H. Schütze, 1999. *Foundations of statistical natural language processing*. MIT press. 25

(Mascaro et al., 2012) C. M. Mascaro, A. Novak, & S. Goggins, 2012. Shepherding and censorship : Discourse management in the tea party patriots facebook group. Dans les actes de *System Science (HICSS), 2012 45th Hawaii International Conference on*, 2563–2572. IEEE. 16

(Mather et Sutherland, 2011) M. Mather & M. R. Sutherland, 2011. Arousal-biased competition in perception and memory. *Perspectives on psychological science* *6*(2), 114–133. 110, 135

(McDonald et al., 2015) G. McDonald, R. Deveaud, R. McCreadie, C. Macdonald, & I. Ounis, 2015. Tweet enrichment for effective dimensions classification in online reputation management. Dans les actes de *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, 654–657. 20, 25, 50, 51, 74, 75, 156

(Meij et al., 2012) E. Meij, W. Weerkamp, & M. de Rijke, 2012. Adding semantics to microblog posts. Dans les actes de *Proceedings of the fifth ACM international conference on Web search and data mining*, 563–572. ACM. 20

(Metaxas et al., 2011) P. T. Metaxas, E. Mustafaraj, & D. Gayo-Avello, 2011. How (not) to predict elections. Dans les actes de *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 165–171. IEEE. 16

(Metzler et Croft, 2005) D. Metzler & W. B. Croft, 2005. A markov random field model for term dependencies. Dans les actes de *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 472–479. ACM. 18

(Metzler et al., 2009) D. Metzler, J. Novak, H. Cui, & S. Reddy, 2009. Building enriched document representations using aggregated anchor text. Dans les actes de *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 219–226. ACM. 20

(Mikolov et al., 2013a) T. Mikolov, K. Chen, G. Corrado, & J. Dean, 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*. 70

(Mikolov et al., 2013b) T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, & J. Dean, 2013b. Distributed representations of words and phrases and their compositionality. Dans les actes de *Advances in Neural Information Processing Systems*, 3111–3119. 47

(Miller, 1995) G. A. Miller, 1995. Wordnet : a lexical database for english. *Communications of the ACM* *38*(11), 39–41. 8, 16

(Mitchell, 1999) T. M. Mitchell, 1999. Machine learning and data mining. *Communications of the ACM* *42*(11), 30–36. 18

(Morstatter et al., 2013) F. Morstatter, J. Pfeffer, H. Liu, & K. M. Carley, 2013. Is the sample good enough ? comparing data from twitter's streaming API with twitter's firehose. Dans E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, & I. Soboroff (Eds.), *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. The AAAI Press. 110

(Mostafa, 2013) M. M. Mostafa, 2013. More than words : Social networks text mining for consumer brand sentiments. *Expert Systems with Applications* *40*(10), 4241–4251. 17

(Navigli, 2009) R. Navigli, 2009. Word sense disambiguation : A survey. *ACM Computing Surveys (CSUR)* *41*(2), 10. 159

(Pan et Yang, 2010) S. J. Pan & Q. Yang, 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* *22*(10), 1345–1359. 18

(Park et al., 2011) S. Park, M. Ko, J. Kim, Y. Liu, & J. Song, 2011. The politics of comments : predicting political orientation of news stories with commenters' sentiment patterns. Dans les actes de *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 113–122. ACM. 16, 79

(Peetz, 2015) H. Peetz, 2015. *Time-Aware Online Reputation Analysis*. doctorat en informatique, Information and language processing systems, University of Amsterdam, Amsterdam, Netherlands. 53, 132, 133

(Peleja et al., 2014) F. Peleja, J. Santos, & J. Magalhães, 2014. Reputation analysis with a ranked sentiment-lexicon. Dans les actes de *Proceedings of the 37th SIGIR conference*. 106

(Pennacchiotti et Popescu, 2011) M. Pennacchiotti & A.-M. Popescu, 2011. A machine learning approach to Twitter user classification. Dans les actes de *ICWSM*, 281–288. 80, 84, 85, 88, 89

(Pupier, 1998) P. Pupier, 1998. Une première systématique des évaluatifs en français. *Revue québécoise de linguistique* *26*(1), 51–78. 159

(Qureshi et al., 2014) M. A. Qureshi, C. Oâ€™Riordan, & G. Pasi, 2014. Exploiting wikipedia for entity name disambiguation in tweets. Dans les actes de *NLP and Information Systems*. 20, 56

(Rahimi et al., 2014) A. Rahimi, M. Sahlgren, A. Kerren, & C. Paradis, 2014. The stavicta group report for replab 2014 reputation dimensions task. Dans les actes de *CLEF*. 50, 51, 74, 75, 156

(Ramírez-de-la Rosa et al., 2014) G. Ramírez-de-la Rosa, E. Villatoro-Tello, H. Jiménez-Salazar, & C. Sánchez-Sánchez, 2014. Towards automatic detection of user influence in Twitter by means of stylistic and behavioral features. Dans les actes de *Human-Inspired Computing and Its Applications*, 245–256. Springer. 88, 89, 95, 98, 123, 126

(Rangel et al., 2015) F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, & W. Daelemans, 2015. Overview of the 3rd author profiling task at PAN 2015. Dans les actes de *Experimental IR meets Multilinguality, Multimodality, and Interaction*. 81, 100, 131

(Rangel et al., 2014) F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, & W. Daelemans, 2014. Overview of the 2nd author profiling task at pan 2014. Dans les actes de *CLEF Evaluation Labs and Workshop*. 45, 80, 100

(Rao et al., 2010) D. Rao, D. Yarowsky, A. Shreevats, & M. Gupta, 2010. Classifying latent user attributes in Twitter. Dans les actes de *CIKM SMUC Workshop*, 37–44. 80, 86, 87, 89, 92

(Riloff et Wiebe, 2003) E. Riloff & J. Wiebe, 2003. Learning extraction patterns for subjective expressions. Dans les actes de *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 105–112. Association for Computational Linguistics. 159

(Roy, 1991) B. Roy, 1991. The outranking approach and the foundations of electre methods. *Theory and decision* *31*(1), 49–73. 28

(Ruthven et Lalmas, 2003) I. Ruthven & M. Lalmas, 2003. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review* *18*(02), 95–145. 20

(Sadikov et al., 2009) E. Sadikov, A. Parameswaran, & P. Venetis, 2009. Blogs as predictors of movie success. 16

(Sadilek et al., 2012) A. Sadilek, H. A. Kautz, & V. Silenzio, 2012. Modeling spread of disease from social interactions. Dans les actes de *ICWSM*. 16, 79

(Sakaki et al., 2010) T. Sakaki, M. Okazaki, & Y. Matsuo, 2010. Earthquake shakes twitter users : real-time event detection by social sensors. Dans les actes de *Proceedings of the 19th international conference on World wide web*, 851–860. ACM. 16, 79

(Saleiro et al., 2013) P. Saleiro, L. Rei, A. Pasquali, C. Soares, J. Teixeira, F. Pinto, M. Nozari, C. Félix, & P. Strecht, 2013. Popstar at replab 2013 : Name ambiguity resolution on twitter. *CLEF 2013 Eval. Labs and Workshop Online Working Notes*. 56, 57, 155

(Salton, 1971) G. Salton, 1971. *The SMART Retrieval System - Experiments un Automatic Document Processing*. Englewood Cliffs. 20, 107

(Salton et Buckley, 1997) G. Salton & C. Buckley, 1997. Improving retrieval performance by relevance feedback. *Readings in information retrieval* *24*(5), 355–363. 20

(Salton et McGill, 1983) G. Salton & M. McGill, 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill. 107

(Salton et al., 1975) G. Salton, A. Wong, & C.-S. Yang, 1975. A vector space model for automatic indexing. *Communications of the ACM* *18*(11), 613–620. 17, 23, 25

(Salton et al., 1974) G. Salton, C.-S. Yang, & C. T. Yu, 1974. A theory of term importance in automatic text analysis. Rapport technique, Cornell University. 18

(Schlesinger et al., 2001) J. D. Schlesinger, D. J. Backer, & R. L. Donway, 2001. Using Document Features and Statistical Modeling to Improve Query-Based Summarization. Dans les actes de *DUC'01*, New Orleans, LA. 107

(Settles, 2012) B. Settles, 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* *6*(1), 1–114. 20, 133

(Sharifi et al., 2010) B. Sharifi, M.-A. Hutton, & J. K. Kalita, 2010. Experiments in microblog summarization. Dans les actes de *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, 49–56. IEEE. 107

(Simon, 1971) H. A. Simon, 1971. Designing organizations for an information-rich world. *Computers, communication, and the public interest 37*, 40–41. 5

(Smeaton, 1999) A. F. Smeaton, 1999. Using nlp or nlp resources for information retrieval tasks. Dans les actes de *Natural language information retrieval*, 99–111. Springer. 18

(Sobkowicz et Sobkowicz, 2012) P. Sobkowicz & A. Sobkowicz, 2012. Properties of social network in an internet political discussion forum. *Advances in Complex Systems* *15*(06), 1250062. 16, 79

(Sparck Jones, 1972) K. Sparck Jones, 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* *28*(1), 11–21. 25

(Sparck-Jones, 2007) K. Sparck-Jones, 2007. Automatic summarising : The state of the art. *Information Processing and Management* *43*(6), 1449–1481. 107

(Spina, 2014) D. Spina, 2014. *Entity-Based Filtering and Topic Detection for Online Reputation Monitoring in Twitter*. doctorat en informatique, Inteligent Systems, UNED, Valencia, Spain. 53, 56

(Spina et al., 2013) D. Spina, J. Carrillo-de Albornoz, T. Martın, E. Amigó, J. Gonzalo, & F. Giner, 2013. Uned online reputation monitoring team at replab 2013. Dans les actes de *CLEF 2013 Eval. Labs and Workshop Online Working Notes*. 60

(Tanaka, 1993) J. S. Tanaka, 1993. Multifaceted conceptions of fit in structural equation models. *Sage focus editions 154*, 10–10. 105, 110

(Tenenhaus et al., 2004) M. Tenenhaus, S. Amato, & V. Esposito Vinzi, 2004. A global goodness-of-fit index for PLS structural equation modelling. Dans les actes de *XLII SIS scientific meeting*, Volume 1, 739–742. 112

(Teufel et Moens, 1997) S. Teufel & M. Moens, 1997. Sentence Extraction as a Classification Task. Dans I. Mani & M. T. Maybury (Eds.), *ACL/EACL97-WS*, Madrid, Spain. 107

(Theodoridis et Koutroumbas, 2009) S. Theodoridis & K. Koutroumbas, 2009. *Pattern Recognition (4th edition)*. Academic Press. 18

(Tombros et al., 1998) A. Tombros, M. Sanderson, & P. Gray, 1998. Advantages of Query Biased Summaries in Information Retrieval. Dans E. Hovy & D. R. Radev (Eds.), *AAAI98-S*, Stanford, California, USA, 34–43. The AAAI Press. 107

(Tommasel et Godoy, 2015) A. Tommasel & D. Godoy, 2015. A novel metric for assessing user influence based on user behaviour. Dans les actes de *SocInf*, 15–21. 92

(Torres-Moreno, 2011a) J. Torres-Moreno, 2011a. *Resume automatique de documents : une approche statistique*. Hermes-Lavoisier. 107

(Torres-Moreno et al., 2005) J. Torres-Moreno, P. Velázquez-Morales, & J. Meunier, 2005. *CORTEX, un algorithme pour la condensation automatique de textes*. Dans les actes de *ARCo*, Volume 2, 365. 115

(Torres-Moreno, 2011b) J.-M. Torres-Moreno, 2011b. Fresa 1.0 (a framework for evaluating summaries automatically). 114

(Torres-Moreno, 2012) J.-M. Torres-Moreno, 2012. Artex is another text summarizer. *arXiv preprint arXiv :1210.3312*. 109, 113

(Torres-Moreno, 2014) J.-M. Torres-Moreno, 2014. Three statistical summarizers at clef-inex 2014 tweet contextualization track. Dans les actes de *4th International Conference of the CLEF initiative*, Sheffield (UK). 69, 70, 110

(Torres-Moreno et al., 2010) J.-M. Torres-Moreno, H. Saggion, I. da Cunha, & E. SanJuan, 2010. Summary Evaluation With and Without References. *Polibits : Research journal on Computer science and computer engineering with applications 42*, 13–19. 114

(Tumasjan et al., 2010) A. Tumasjan, T. O. Sprenger, P. G. Sandner, & I. M. Welpe, 2010. Predicting elections with twitter : What 140 characters reveal about political sentiment. *ICWSM 10*, 178–185. 16

(Uddin et al., 2014) M. M. Uddin, M. Imran, & H. Sajjad, 2014. Understanding types of users on Twitter. *arXiv cs.SI*, 1406.1335. 80, 87, 89, 92

(Vapnik et Vapnik, 1998) V. N. Vapnik & V. Vapnik, 1998. *Statistical learning theory*, Volume 1. Wiley New York. 25

(Velcin et al., 2014) J. Velcin, Y. Kim, C. Brun, J. Dormagen, E. SanJuan, L. Khouas, A. Peradotto, S. Bonnevay, C. Roux, J. Boyadjian, et al., 2014. Investigating the image of entities in social media : Dataset design and first results. Dans les actes de *LREC*. 37, 105

(Vilares et al., 2014) D. Vilares, M. Hermo, M. A. Alonso, C. Gómez-Rodrıguez, & J. Vilares, 2014. Lys at clef replab 2014 : Creating the state of the art in author influence ranking and reputation classification on twitter. Dans les actes de *CLEF 2014*, 1468–1478. 85, 89, 95, 96

(Villatoro-Tello et al., 2014) E. Villatoro-Tello, G. Ramirez-de-la Rosa, C. Sanchez-Sanchez, H. Jiménez-Salazar, W. A. Luna-Ramirez, & C. Rodriguez-Lucatero, 2014. Uamclyr at replab 2014 : Author profiling task. Dans les actes de *4th International Conference of the CLEF initiative*. 95, 96

(Villegas, 2013) A. M. Villegas, 2013. *Compression automatique de phrases : une étude vers la génération de résumés*. doctorat en informatique, Laboratoire Informatique d'Avignon, Université d'Avignon et des Pays de Vaucluse, Avignon, France. 107

(Villena Román et al., 2013) J. Villena Román, S. Lana Serrano, E. Martínez Cámara, & J. C. González Cristóbal, 2013. Tass-workshop on sentiment analysis at sepln. 36, 105

(Wang, 2010) A. H. Wang, 2010. Don't follow me : Spam detection in Twitter. Dans les actes de *International Conference on Security and Cryptography*, 1–10. 79, 86, 88, 90, 92

(Weerkamp et al., 2009) W. Weerkamp, K. Balog, & M. de Rijke, 2009. A generative blog post retrieval model that uses query expansion based on external collections. Dans les actes de *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2-Volume 2*, 1057–1065. Association for Computational Linguistics. 20

(Weerkamp et al., 2011) W. Weerkamp, S. Carter, & M. Tsagkias, 2011. How people use twitter in different languages. 58

(Wen et Marshall, 2014) D. Wen & G. Marshall, 2014. Automatic twitter topic summarization. Dans les actes de *Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on*, 207–212. IEEE. 108

(Weng et al., 2010) J. Weng, E.-P. Lim, J. Jiang, & Q. He, 2010. TwitterRank : finding topic-sensitive influential twitterers. Dans les actes de *WSDM*, 261–270. 80, 86, 88

(Weren et al., 2014) E. R. D. Weren, A. U. Kauer, L. Mizusaki, V. P. Moreira, J. P. M. de Oliveira, & L. K. Wives, 2014. Examining multiple features for author profiling. *Journal of Information and Data Management* *5*(3), 266. 88, 89

(Wold, 1982) H. Wold, 1982. Soft modeling : the basic design and some extensions. Dans les actes de *Systems under indirect observations : Causality, structure, prediction*, 36–37. North-Holland. 111

(Zhai et Lafferty, 2001a) C. Zhai & J. Lafferty, 2001a. The dual role of smoothing in the language modeling approach. Dans les actes de *Proceedings of the Workshop on Language Models for Information Retrieval (LMIR) 2001*. Citeseer. 26

(Zhai et Lafferty, 2001b) C. Zhai & J. Lafferty, 2001b. A study of smoothing methods for language models applied to ad hoc information retrieval. Dans les actes de *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 334–342. ACM. 26

(Zhai et Lafferty, 2004) C. Zhai & J. Lafferty, 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)* *22*(2), 179–214. 26

(Zhao et al., 2011) W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, & X. Li, 2011. Topical keyphrase extraction from twitter. Dans les actes de *Proceedings of the 49th Annual Meeting of the ACL : Human Language Technologies*. 106

(Zhu, 2005) X. Zhu, 2005. Semi-supervised learning literature survey. 19

(Zingla M-A. et Y., 2014) L. C. Zingla M-A., Ettaleb M. & S. Y., 2014. Inex2014 : Tweet contextualization using association rules between terms. Dans les actes de *4th International Conference of the CLEF initiative*, Sheffield (UK). 69, 70

(Zubiaga et al., 2012) A. Zubiaga, D. Spina, E. Amigó, & J. Gonzalo, 2012. Towards real-time summarization of scheduled events from twitter streams. Dans les actes de *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, 319–320. 107, 108

Annexe A

Participations aux campagnes RepLab 2013-2014

Approche	F-Score	Précision	F(R,S)
(Cossu et al., 2013) *Lia_Prio_5*	**.571**	.636	**.335**
CRF	.554	.633	.318
CRF w/ Context	.551	.631	.318
SVM w/ Context	.564	.645	.304
SVM	.563	.644	.304
(Cossu et al., 2013) *KBA*	.421	.585	.282
Baseline	.512	.570	.274
Cosine w/ Context	.562	.634	.260
Cosine	.561	.633	.260
(Cossu et al., 2013) *Fusion PROMETHEE*	-	.651	.253
(Cossu et al., 2013) *Combinaison Linéaire*	-	.647	.251
(Cossu et al., 2013) *Fusion ELECTRE*	-	**.652**	.251
Median	-	*.573*	*.249*

TABLE *A.1* – *Performances des systèmes de détection de priorité sur les données d'évaluation de la collection RepLab* 2013 *ordonnées selon la métrique officielle du challenge : F(R,S).*

Approche	Précision	F(R,S)
(Saleiro et al., 2013)	.908	**.488**
(Hangya et Farkas)	**.928**	.438
(Cossu et al., 2013) *Lia_Filter_1*	.872	.381
(Cossu et al., 2013) *Lia_Filter_6*	.876	.341
Baseline	*.871*	*.325*
(Cossu et al., 2013) *Fusion PROMETHEE*	.882	.312
(Cossu et al., 2013) *Fusion ELECTRE*	.879	.302
(Cossu et al., 2013) *Combinaison Linéaire*	.874	.296
(Cossu et al., 2013) *KBA*	.850	.289
Cosine	.835	.272
Median	*.826*	*.265*
(Cossu et al., 2013) *Lia_Filter_7*	.806	.187
(Cossu et al., 2013) *Lia_Filter_3*	.800	.126

TABLE *A.2* – *Performances des systèmes de filtrage sur les données d'évaluation de la collection RepLab* 2013 *ordonnées selon la métrique officielle du challenge : F(R,S).*

Approche	F-Score (-U)	Précision (-U)	F-Score (+U)	Précision (+U)
CRF w/ Context	.492	**.771***	.481	.761
CRF	.491	.769*	.483	.762
Cosine w/ Context	**.505**	.739	.494	.707
Cosine	.491	.736	.500	.693
SVM	.469	.732	.461	.679
SVM w/ Context	.468	.732	.456	.679
(McDonald et al., 2015)	.473	.731	-	-
(Rahimi et al., 2014)	.489	.695	-	-
SVM Baseline	.38	.622	-	-
Naive Baseline	.152	.560	-	-

TABLE *A.3* – *Performances des systèmes automatique de catégorisation thématiques sur les données d'évaluation de la collection RepLab* 2014 *ordonnés par Précision (-U). Les méthodes utilisant la généralisation lexicale sont notées (w/ Context). Les scores les plus élevés sont indiqués en gras.*

Annexe B

Expérimentations avec les données Vodkaster

Cette annexe présente des expérimentations complémentaires menées en début de thèse durant les phases de collectes et d'annotations des données Imagiweb. Ces expérimentations avaient pour objectif d'évaluer la pertinence des méthodes que nous avons présenté dans le chapitre 2 puis utilisé dans le chapitre 4 sur différentes collections de données de nature similaire. Nous commençons cette annexe par la description de la collection de données que nous avons utilisé à cet effet, cette dernière provient du portail communautaire Vodkaster[1] Nous présentons ensuite les résultats de notre méthode de catégorisation de contenus textuels que nous analysons. Ces travaux ont fait l'objet de publications (voir (Cossu et al., 2013,?)).

Introduction et Problématiques

B.0.1 Introduction

La plateforme Vodkaster a été développée sur un concept de diffusion d'extraits de films. Son site permet de découvrir des films via des extraits (généralement les scènes les plus marquantes ou célèbres d'un film). Cependant Vodkaster s'est démarqué en permettant aux utilisateurs de participer à l'enrichissement de la base d'extraits, mais également de critiquer les oeuvres. Ces critiques prennent la forme de messages courts (appelés « *micro-critiques* » ou μC) équivalents aux Micro-Blogs de Twitter. Nous souhaitons à partir du contenu

1. www.vodkaster.com

textuel de ces critiques être en mesure de retrouver l'opinion complète de l'utilisateur (la tonalité de la critique et l'objet sur lequel elle porte). Ce problème de catégorisation se matérialise par la prédiction d'une tranche de notes (positive, neutre ou négative) qu'un utilisateur du site Vodkaster a attribué à un film. La prédiction est tout d'abord faite uniquement à partir du contenu de à la « *micro-critique* » associé au couple (utilisateur, film). Enfin, nous considérons un critère temporel pour sélectionner les données qui permettraient d'améliorer la qualité de nos prédictions. Contrairement aux autres systèmes de recommandation ou de prédiction de notes existants, ici la note n'est plus l'élément central. Néanmoins, sa présence permet, lors de la phase d'apprentissage, de pouvoir affecter un score (positif, neutre ou négatif) à un terme (ou une chaîne) porteur d'opinion. La note devient le point d'appui de l'extraction du contenu de la critique à laquelle elle est associée. Cette tâche est loin d'être évidente lorsque l'on se retrouve face à des exemples tels que l'avis de l'utilisateur « *peonidelavega* » au sujet du film « *Le Pacte* » : « *Chef d'oeuvre ! non je plaisante...* » (sa note est de $0,5$).

Apports des méta-données et sur-apprentissage

Dans l'optique d'alimenter un système de recommandation basé sur le contenu des avis des utilisateurs, nous développons un système permettant, grâce à l'apprentissage automatique, de prédire la catégorie de note d'une critique. Notre travail s'articule de la manière suivante : premièrement, nous nous basons sur la répartition des critiques en fonction de leur catégorie dans le but d'extraire les termes (ou chaînes de mots) relatives à chacune des catégories d'opinions. Ces chaînes ainsi extraites sont ensuite utilisées dans un système automatique de catégorisation supervisée de textes. Toutefois, les opinions catégorisées sont considérées globalement sans prise en compte de l'identité de l'utilisateur ou du titre du film. De plus, il arrive qu'avec le temps l'utilisateur change d'avis, deviennent plus ou moins sévère mais également que l'opinion générale au sujet d'un film évolue laissant entendre qu'il faut être capable de sélectionner à l'instant t pour un film et un utilisateur les données qui correspondent le mieux aux évolutions d'opinions les plus récentes.

Détection semi-automatique d'aspects

Si il est clair que les mots porteurs d'opinions jouent un rôle important pour déterminer l'orientation du message, identifier les cibles auxquelles les messages se rapportent pour en contextualiser la portée reste une tâche difficile. L'analyse peut également être menée dans l'autre sens, lorsque en cherchant

dans le contexte d'une cible, on souhaite détecter les termes les plus polarisés. Un des objectifs visés est l'extraction de couples (cible, marqueur de polarité) permettant à la fois de catégoriser le message mais également de constituer un résumé de la représentation de l'entité « *film* ». A l'inverse des expériences présentées dans cette thèse, nous ne disposons pas de cibles pré-établies, celles que nous seront en mesure de détecter ne seront pas limités aux seuls concepts identifiables par des experts du domaine [2], mais doivent émerger des avis analysés conformément à la façon dont ils ont été exprimés. Cette façon de procéder tient implicitement compte de la restriction des différents sens d'un mot à ceux qui ont cours dans le domaine abordé par les auteurs des critiques (Riloff et Wiebe, 2003). Citons par exemple le cas du terme « *navet* » qui est un légume plus ou moins apprécié par les gastronomes mais aussi et surtout pour ce qui nous concerne un mauvais film dans le domaine du cinéma. Il serait possible de se baser sur des listes de marqueurs d'opinions comme le propose (Navigli, 2009) mais il serait nécessaire d'établir leur polarité dans le contexte de la collection de données ce qui impliquerait une coûteuse opération de désambiguïsation lexicale.

B.1 Cadre expérimental

Collection de données

La collection de données Vodkaster contient 77,000 μC, 20,000 contributions sont utilisées pour constituer les corpus de développement et test (10,000 chacun), l'ensemble des μC restantes composent l'ensemble d'entraînement. Ce découpage a été fait tout en respectant l'ordre chronologique des critiques afin d'éviter un entre-laçage [3]. Dans le cadre de nos expériences, l'échelle des notes est de façon volontaire réduite aux trois barreaux habituellement considérés dans la littérature (positif, neutre et négatif). Les seuils des barreaux ont été déterminés de façon empirique : positif (note supérieure ou égale à quatre) et négatif (note inférieure ou égale à deux).

2. Comme dans le cadre du projet Imagiweb ou par ce qui est communément admis (Pupier, 1998). A notre connaissance, au moment où nous avons mené ces expériences, une telle ressources (liste de cibles potentielles sur des oeuvres cinématographiques) n'existe tout simplement pas en français.

3. L'entre-laçage est un phénomène qui apparaît lorsque l'on rencontre dans les tests des éléments déjà appréciés au moment de l'apprentissage. Ce phénomène biaise les résultats pour des données où les utilisateurs se répondent les uns aux autres sur un sujet donné. Tout sujet croisé dans l'apprentissage qui est recroisé dans le test est de fait reconnu et cela facilite de façon artificielle la tâche de classification. La notion temporelle fera l'objet d'explications et expérimentations supplémentaires par la suite

Malgré les tailles restreintes des critiques et la liste de cible, les utilisateurs arrivent à exprimer plusieurs opinions (parfois opposées) sur les différents éléments des films. Nous avons tablé sur le fait que les positions les plus tranchées feraient ressortir plus de cibles associées à des qualificatifs que les critiques plus nuancées. Les critiques nuancées ou équilibrées (μC contenant un des «*pivots*» prédéterminés) sont retirées des différents corpus pour nos expérimentations. Nous avons à cet effet sélectionné uniquement les deux «*pivots*» les plus fréquents [4] dans le corpus d'apprentissage : «*mais*» et «*malgré*». Ne seront donc présentes dans les collections de développement et de test que les μC *a priori* fortement polarisées contenant au moins une cible et ne contenant aucun des «*pivots*» de langage retenus. Ces contraintes réduisent les sous-ensembles de développement et évaluation à 5,010 critiques sur l'ensemble des 10,000 présentes à l'origine. A titre informatif, voici quelques caractéristiques au sujet de la collection de données :

— Il y a 1,723 contributeurs différents soit en moyenne 24 critiques par contributeur ;
— 510 utilisateurs n'ont émis qu'une seule critique ;
— Le plus gros contributeur a posté 1559 critiques depuis son inscription à l'ouverture de la plateforme ;
— 7,415 films ont été critiqués soit en moyenne 5 critiques par film ;
— 3,400 films n'ont été critiqués qu'une seule fois ;
— Le film (Drive) le plus critiqué l'a été 241 fois depuis sa sortie ;

Le tableau B.1 illustre un exemple de ce problème pour un des films les plus critiqués du corpus : X-Men. Des termes *a priori* élogieux, sont utilisés avec une grande variabilité de note :

terme	moyenne	min	max	fréquence
meilleur	4,05	3	5	9
bon	3,55	2	5	9
bien	3,11	1	4	9
très-bon	3	2	4	3

TABLE *B.1* – *Dispersion des termes «positifs» pour le film X-Men.*

4. On aurait pu en rajouter d'autres comme «*bien que*» et «*et pourtant*» (130 et 150 occurrences).

Propositions

Métadonnées et apprentissage

Nous disposons avec le corpus de données de nombreuses informations supplémentaires concernant la critique :

— l'identité de l'auteur de la critique (son pseudo sur le portail) ;
— le titre du film dont il est question dans la critique ;
— la date d'émission de la critique.

Afin de tenir compte du comportement passif de l'utilisateur mais également de celui du film, nous avons décidé d'inclure les deux premières données dans le système. Cela se matérialise par la prise en compte dans la critique du pseudo et du titre du film en les intégrant tels quels dans le sac de mot la critique au même titre que les termes qui la compose. De lui même, par le jeu des pondérations statistiques, le système va intégrer dans le modèle les éléments rajoutés. Nous testons également des combinaisons temporelles en utilisant dans notre corpus de test des critiques plus récentes que celles utilisées dans le corpus d'apprentissage. Même si globalement prédire la note d'une critique ancienne à partir de critiques récentes ne parait pas pertinent, la prédiction pour des critiques récentes est différente selon si l'on prend en compte des critiques proches ou lointaines. Ce constat nous amène à réfléchir à la nécessité d'adapter nos modèles aux évolutions du jugement de l'utilisateur et aux mouvements d'opinions sur un sujet donné.

Détection et extraction d'aspects

Nous avons considéré une première étape qualifiée d'apprentissage, qui à partir des données permet d'obtenir automatiquement les marqueurs de polarité les plus importants. A partir de cette base, nous cherchons de manière automatique les cibles qui apparaissent le plus fréquemment à proximité de ces marqueurs d'opinions. Ensuite, nous construisons un ensemble de couples (marqueur de polarité, cible) pour montrer qu'en s'appuyant sur ces couples, on arrive à expliquer plus finement les prises de positions tout en maintenant (voire améliorant) le niveau de performance du système de catégorisation automatique. Deux systèmes concurrents ont été mis en place : l'un prenant en compte le couple (cible-marqueur de polarité), l'autre se basant sur l'ensemble des termes présents dans la μC. Toutes les expériences présentées tiennent compte éléments supplémentaires décrits précédemment. C'est à dire (i) la polarité du pseudo de l'utilisateur et (ii) celle portée par du titre du film. Ces éléments porteurs d'opinions ont été intégrés comme des termes à l'intérieur de la μC. Notre méthode consiste à extraire dans le corpus les éléments les

plus porteurs d'opinions (marqueurs de polarité). Une fois ceux-ci extraits, nous cherchons, à proximité de ces derniers, s'il existe des éléments à pouvoir discriminant plus modéré, non présents dans un anti-dictionnaire (composé principalement de mots-outils). Si la fréquence de ces éléments dépasse un plancher déterminé empiriquement, nous pouvons les considérer comme des « *cibles* » sur lesquelles portent les opinions. Par la suite, nous pouvons considérer l'ensemble de ces cibles (au même titre que des métadonnées film ou pseudo) comme des Entités Nommées Conceptuelles (ENC).

Puis, l'enrichissement de la liste de cibles se fait selon les deux procédures suivantes :

— Procédure P1 : trouver des cibles permettant de couvrir des μC où aucun terme n'appartient à la liste de cibles. Ne sont alors retenus que les termes présents dans le plus grand nombre de μC *résiduelles* mais qui permettraient également d'améliorer la couverture des μC déjà sélectionnées. Les termes ayant un pouvoir discriminant faible (seuil fixé empiriquement) sont filtrés.

— Procédure P2 : dans le cas de μC correctement étiquetées par M1 mais pas par M2. L'objectif est de chercher dans le voisinage du terme de polarité P, qui a le plus contribué à la bonne décision de M1, un terme T n'étant ni dans la liste des cibles ni dans un anti-dictionnaire. Seront alors proposés les termes se trouvant dans le plus grand nombre de μC résiduelles, avec fréquence élevée et pouvoir discriminant important (seuil fixé empiriquement).

Itérer ces deux procédures a permis d'augmenter facilement la couverture de la liste de cibles en refrénant l'accroissement de sa taille (passant de 550 à 982 cibles). Parmi les $5,010$ μC restantes dans le corpus après retrait de celles contenant un « *pivot* », $4,580$ contiennent *au moins* une des cibles présentes dans la liste. La couverture est d'environ $2,9$ cibles par μC traitée avec 550 cibles. En s'appuyant sur les marqueurs de polarité se trouvant à proximité des cibles, et donc en filtrant ce que l'on peut considérer comme du bruit, on cherche à éliminer une partie de ce qui pourrait amener à prendre une mauvaise décision.

Expérimentations et résultats

Intérêt des métadonnées

L'ajout de métadonnées dans la critique semble avoir une influence positive sur l'opération de catégorisation bien que non significative avec un gain absolu de trois points en termes de F-Score (Cossu et al., 2013). Toutefois, comme nous avons pu le constater, toutes les métadonnées ne se valent pas car l'ajout du titre

du film offre une amélioration plus significative que l'ajout du pseudo. Ce qui amène au constat suivant, le système de classification semble avoir un problème bien connu dans le domaine de la recommandation de contenu : le démarrage à froid. Ici, cela se traduit par de mauvais résultats sur de nouveaux films. A ce titre, l'analyse de nos 849 erreurs nous permet d'obtenir plus d'informations :

Catégorie positive	Nombre d'erreurs	Taux d'erreurs (en %)
posteur inconnu film connu	105	12,36
posteur connu film inconnu	112	13,19
film inconnu posteur inconnu	40	4,71
film connu posteur connu	143	16,84
Catégorie négative		
posteur inconnu film connu	76	8,95
posteur connu film inconnu	179	21,08
film inconnu posteur inconnu	34	4
film connu posteur connu	160	18,8

TABLE *B.2* – *Taux d'erreurs*

Le tableau B.2 montre que le taux d'erreurs est plus important pour les critiques dont le film n'a jamais été croisé dans l'apprentissage. Dans ce cas, l'apport de la connaissance de l'utilisateur émettant la critique n'aide pas le système.

Quantités de données d'apprentissage

Comme évoqué précédemment, prédire la note d'une critique ancienne à partir de critiques récentes ne parait pas pertinent. Toutefois, la prédiction pour des critiques récentes avec des critiques anciennes présente un paradoxe. En effet au delà d'un certain seuil, augmenter la masse d'apprentissage par retour dans le passé n'améliore pas les performances. L'influence observée est parfois même négative. A taille égale les données d'apprentissage les plus proches dans le temps des données de test offrent même de meilleurs résultat pour une période de test donnée comme le montre la figure B.1 :

D'autres tests ont été menés afin de détecter un taille ou segment d'apprentissage optimal. Les tests ont montré que pour chaque sous corpus de test il existe un sous corpus d'apprentissage qui donne de meilleurs résultats que le corpus d'apprentissage considéré dans son ensemble. Il convient toutefois de relativiser, les critiques de cinéma sont liées, voire influencées, par des phénomènes d'actualité. On observe notamment des pics d'activités durant les festivals, morosité après une élection ou faits divers tragiques. Les données ne

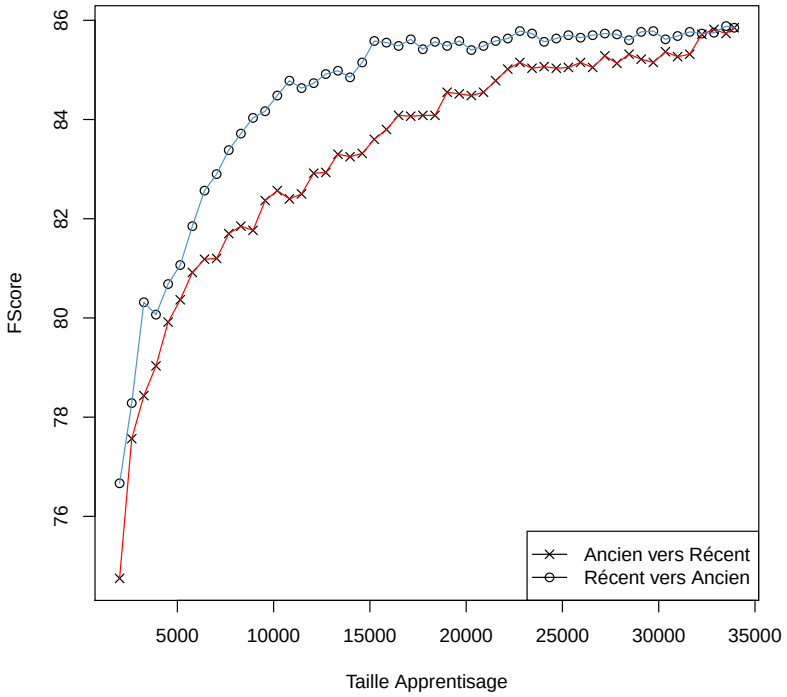


FIGURE *B.1* – *Évolution du F-Score en fonction de la taille d'apprentissage.*

disposent pas d'une profondeur temporelle suffisante pour asseoir de manière fiable ces résultats.

Utilisation des aspects

Au lieu d'opter pour un protocole lourd d'évaluation de la pertinence des cibles détectées nous avons choisi d'en faire une estimation certes grossière mais peu coûteuse. Leur extraction peut être considérée comme valide dès que la prise en compte des seuls couples présents permet de faire aussi bien qu'un système de catégorisation utilisant l'intégralité des termes de la μC comme nous le montrons dans la section suivante où une première série d'expériences a été menée (avec 550 cibles). Nous proposons de comparer un système de catégorisation classique qui se baserait sur l'ensemble du contenu textuel avec une variante qui ne prendrait cette fois en compte que les couples (cible, marqueur de polarité). Les marqueurs de polarité seront recherchés avec un rayon de **R** (variable entre 1 et 9) termes de part et d'autre de la cible. Nous avons fait varier le rayon afin d'évaluer l'impact du contexte sur la catégorisation de la cible.

Les performances présentés sur la figure B.2 sont mesurées en termes de précision [5]. Il arrive parfois pour des petits rayons qu'il n'y ait aucun couple présent dans une μC pour cette raison nous comparons M1 et M2 sur la précision à un même niveau de rappel, celui déterminé par M2(Cossu et al., 2013).

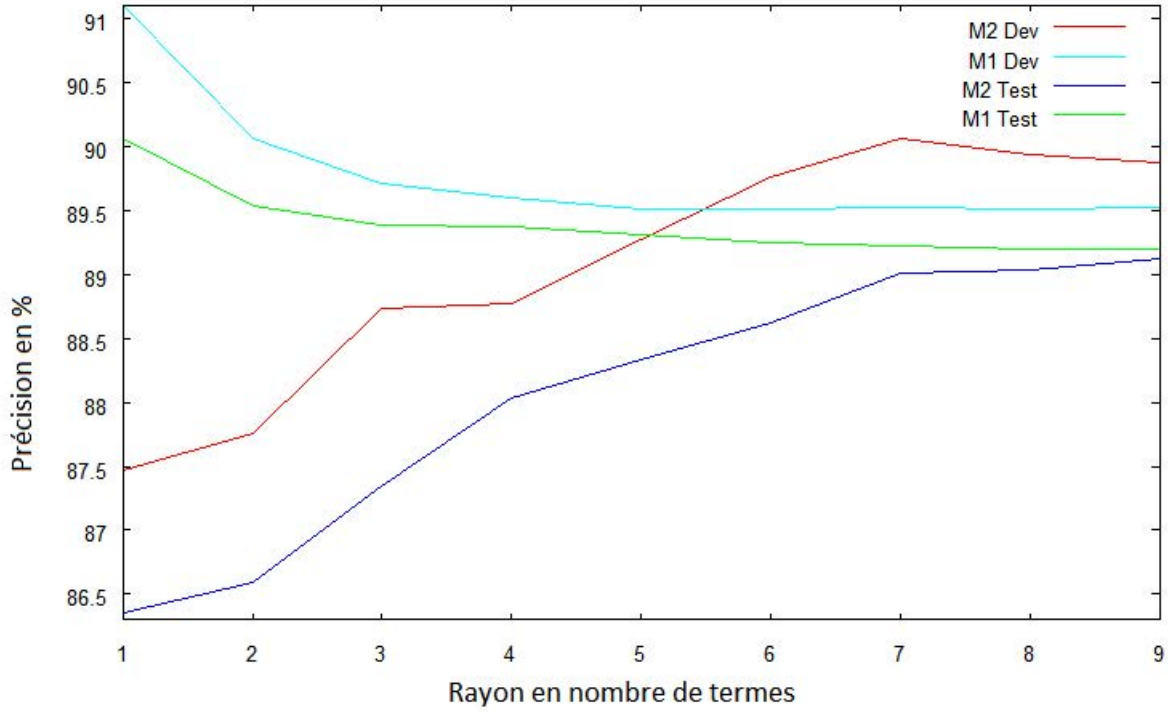


FIGURE *B.2* – *Évolution des résultats en fonction du rayon.*

Avec un rayon égal à 7, on trouvait 4,449 critiques du développement contenant au moins une cible, M1 en classait correctement 3,957 (soit un F-Score de 0,889) contre 3,975 (0,893) pour M2. En ramenant le rayon à 1, il ne restait que 3,286 μC, M1 retrouve correctement la catégorie de notes de 2,977 μC (soit un F-Score de 0,905) contre 2,827 (ou 0,860) pour M2. En réduisant le rayon de la fenêtre dans laquelle, autour d'une cible, sont pris en compte des marqueurs de polarité, les résultats de M2 et notamment le rappel chutent logiquement là où les performances de M1, qui prend en compte l'ensemble du contenu de chaque μC, restent stables. Notons toutefois que M2 permet d'identifier des contenus pour lesquels M1 est bien plus performant que sur l'ensemble du corpus (F-Score de 0,895 sur l'ensemble de développement et 0,889 sur celui d'évaluation). Cette mesure permet donc de faire ainsi un premier filtrage des données à tester. Afin de pouvoir intégrer dans le test des critiques ne contenant pas de cibles (ce qui est le cas pour les critiques très courtes ne contenant parfois qu'un seul terme très souvent porteur de polarité) nous avons considéré que le

5. Valeurs équivalentes au F-Score comme le rappel est fixe, voir section 3.2.

film pouvait être une cible. Ce qui nous permet d'arriver à 982 « *cibles* » potentielles. La couverture passe à environ 3, 4 cibles par μC traitée contre 2, 9 avec la première liste. Nous constatons alors que le passage de 550 à 982cibles permet d'améliorer les résultats et il ne serait pas improbable que les résultats s'améliorent encore avec une liste de cibles plus grande. Pour illustrer notre propos nous donnons quelques exemples (extraits à partir de la première liste de 550 cibles) avec leur nombre d'occurrences sur l'ensemble du corpus : « *acteurs* » (3, 000), « *mise en scène* » (2, 000), « *réalisation* » (931), « *esthétique* »(630).

Conclusion

Nous avons présenté dans cette annexe une approche permettant d'extraire des cibles en fonction d'une liste constituée de manière semi-automatique. Ces cibles sont ensuite utilisées dans une tâche de catégorisation de contenus textuels. Les résultats de nos expériences suggèrent que l'extraction d'un couple peut suffire à catégoriser un contenu textuel de type μC, l'utilisation de ces cibles a permis d'améliorer globalement le finesse du système de catégorisation. Nous avons également examiné l'influence des différentes métadonnées (identité de l'emetteur d'un avis, identifiant du produit critiqué). Il apparaît que pour déterminer plus précisément l'opinion d'un utilisateur sur un produit, il est préférable de considérer son jugement moyen ainsi que le jugement moyen porté sur le produit. De plus, nous montrons qu'avec un nombre égal de documents, l'ensemble d'apprentissage le plus proche (temporellement parlant) des données d'évaluation se montre plus fiable et donnait de meilleurs résultats que les documents plus éloignés dans le temps.