Analyse de sentiments automatique, adaptative et applicative

—

résumé de thèse

Alexander Pak
Université Paris-Sud,
Lab. LIMSI-CNRS, Bâtiment 508,
F-91405 Orsay Cedex, France
alexpak@limsi.fr
traduction française: P. Paroubek

Résumé. L'analyse de sentiments est un des nouveaux défis apparus en traitement automatique des langues avec l'avènement des réseaux sociaux sur le WEB. Profitant de la quantité d'information maintenant disponible, la recherche et l'industrie se sont mises en quête de moyens pour analyser automatiquement les opinions exprimées dans les textes. Dans cet ouvrage, nous nous plaçons dans un contexte multilingue et multi-domaine pour explorer la classification automatique et adaptative de polarité. Plus particulièrement, nous proposons dans un premier temps de répondre au manque de ressources lexicales par une méthode de construction automatique de lexiques affectifs multilingues à partir de microblogs. Nous proposons ensuite, pour une meilleure analyse, de remplacer le traditionnel modèle n-gramme par une représentation à base d'arbres de dépendances syntaxiques. Finalement, nous étudions l'impact que les traits spécifiques aux entités nommées ont sur la classification des opinions minoritaires et proposons une méthode de normalisation des décomptes d'observables, qui améliore la classification de ce type d'opinion. Nos propositions ont été évaluées quantitativement pour différents domaines d'applications (les films, les revues de produits commerciaux, les nouvelles et les blogs) et pour plusieurs langues (anglais, français, russe, espagnol et chinois), avec en particulier une participation officielle à plusieurs campagnes d'évaluation internationales (SemEval 2010, ROMIP 2011, I2B2 2011).

1 État de l'art

La première partie de la thèse présente l'état de l'art en fouille d'opinion et analyse de sentiments. L'analyse de sentiments est un domaine récent en traitement automatique des langues, qui confronte les chercheurs à toute la complexité de la langue naturelle. Si le but recherché en traitement automatique des langues est d'être capable de traiter avec des ordinateurs le langage humain, l'objectif en analyse de sentiments et de pourvoir reconnaître les émotions humaines, telles qu'elles sont exprimées dans les textes. Dans le chapitre 1 nous exposons le thème de notre recherche ainsi que le plan de notre thèse. Dans le chapitre 2, nous définissons les concepts clés à la base de nos travaux. Les termes *analyse de sentiments* (sentiment analysis) et *fouille d'opinion* (opinion mining) sont souvent utilisés de manière interchangeable. D'après [6], le vocable analyse de sentiments est préféré dans le domaine du traitement automatique des langues, tandis que le terme fouille d'opinion a été lui adopté par la communauté de la recherche d'information. Bien que ces deux termes concernent des champs d'investigation très proches, qui pourraient même être considérés comme une seule et même entité, nous avons choisi pour nos travaux d'utiliser le terme analyse de sentiments, que nous distinguons de la fouille d'opinion.

Nous postulons que l'**opinion** est l'expression d'un individu à propos d'un objet ou d'un sujet particulier. Nous qualifions la personne qui s'exprime comme le **porteur d'opinion** (opinion holder) et le sujet de l'expression comme la **cible de l'opinion** (opinion target). Ainsi le terme **fouille d'opinion** se réfère au champ du traitement automatique des langues qui étudie les opinions. Nous distinguons les opinions des

faits, qui sont des informations avérées, comme le sont en particulier les informations que l'on désigne par le terme sens commun. Notre définition de l'opinion, nécessite que lui soient associés un porteur et une déclaration de ce dernier, précisant sa position par rapport à la cible, sinon ce n'est pas une opinion. Par exemple, l'énoncé « j'ai froid » ("I am cold"), n'est pas, d'après nous, une expression d'opinion, car il n'y a pas à proprement parler de positionnement exprimé par rapport à un objet ou un sujet, mais plutôt l'expression d'un fait. Par contre, l'énoncé « j'ai l'impression qu'il fait froid dans cette pièce » est une expression d'opinion, avec comme cible la température de la pièce. De la même manière, l'énoncé « l'économie est en récession » ("Economy is in recession") n'est pas une expression d'opinion, car le porteur n'est pas mentionné explicitement, mais l'énoncé « Le ministre croit que l'économie est en récession » ("The minister believes that the economy is in recession") est une expression d'opinion, avec « le ministre » comme porteur de l'opinion.

Nous définissons le **sentiment** (sentiment) comme le jugement que porte un individu sur un objet ou un sujet, ce jugement étant caractérisé par une **polarité** (polarity) et une **intensité** (intensity). L'**analyse de sentiments** (sentiment analysis) est le champs du traitement automatique des langues qui étudie les sentiments. Pour nous, une polarité est soit positive, soit négative, soit un mélange de ces deux valeurs, tandis que l'intensité montre le degré de positivité ou de négativité, et varie de faible à forte. De notre définition, il ressort qu'un **sentiment est un type particulier d'opinion dotée d'une polarité**. Ainsi nous opposons les sentiments aux faits et aux expressions de neutralité face à un objet ou un sujet particulier.

Par **action bénéfique** (beneficial action), nous entendons une action qui profite au possesseur de la cible de l'opinion. Par exemple, dans le cas des critiques de film, la cible de l'opinion sera un film particulier, le possesseur de la cible sera une compagnie cinématographique et l'action bénéfique sera l'achat d'un billet pour une séance ou l'achat d'un DVD. En politique, la cible de l'opinion pourra être un candidat à une élection et l'action bénéfique, un vote pour ce candidat[1]. Ainsi, **la polarité d'un sentiment sera dite positive si l'opinion est en faveur de l'action bénéfique et elle sera dite négative si elle s'y oppose**. L'intensité d'un sentiment mesure dans ce cas le degré de soutien ou d'opposition à l'action bénéfique. Notons, que le soutien ou l'opposition n'ont pas besoin d'être explicites. Par exemple, écrire une bonne critique pour un film, n'implique pas nécessairement d'inciter explicitement le lecteur à aller voir le film ou à acheter le DVD; le contenu positif de la critique étant une motivation suffisante en soi, qui va susciter, par voie de conséquence, l'achat du film.

Dans le chapitre 3, nous passons en revue les tâches communément effectuées en fouille d'opinion et analyse de sentiments : analyser la subjectivité, détecter les opinions, classer selon la polarité, identifier le porteur et la cible des expressions d'opinion, résumer les opinions, détecter l'ironie, détecter les « fausses » opinions (spams).

Le chapitre 4 est consacré au thème central de nos travaux de thèse : **le classement en polarité** (polarity classification); nous y présentons en détails la problématique scientifique qui concerne essentiellement l'analyse du discours, le traitement des négations, le traitement des métaphores, l'adaptation au domaine et le multilinguisme. Dans ce chapitre, nous faisons aussi un tour d'horizon des données expérimentales et des cadres évaluatifs qui existent pour les algorithmes de classement polaire.

Le chapitre 5 présente les approches existantes pour le classement en polarité, en distinguant les deux grand courants qui sont d'une part les méthodes à base de lexique et d'autre part les méthodes statistiques. Les premières utilisent un lexique affectif pour déterminer la polarité d'un texte, tandis que les secondes mettent en œuvre l'apprentissage automatique sur des textes de polarité connue pour construire des modèles de reconnaissance de cette polarité. Les méthodes à base de lexique sont coûteuses car elles nécessitent un gros travail de la part d'experts pour construire le lexique. A l'opposé les méthodes statistiques sont beaucoup plus faciles à mettre en œuvre mais donnent en général de moins bons résultats, avec la réserve que la qualité des performances augmente avec l'augmentation de la taille des données d'apprentissage.

[1] Dans ce cas, la cible est aussi le possesseur.

2 Nos travaux

La seconde partie de la thèse est dédiée à la présentation de notre contribution à l'analyse de sentiments. Nous avons concentré nos efforts sur les deux dernières problématiques scientifiques présentées dans le chapitre 4, à savoir : l'adaptation au domaine et le multilinguisme. Nous ne voulons pas dépendre d'une ressource spécifique à un domaine particulier, comme par exemple une ontologie « métier ». Nous voulons aussi être, autant que faire se peut, indépendant de la langue, en proposant des algorithmes facilement transportable vers de nouvelles langues cibles. C'est pourquoi nous trouvons au cœur de nos travaux, un classifieur à base d'apprentissage automatique, qui n'a besoin que de données d'apprentissage dans la langue cible. Nous faisons en effet l'hypothèse, qu'il est beaucoup plus facile de collecter du matériaux d'apprentissage dans différentes langues, plutôt que de porter des ressources lexicales vers une nouvelle langue cible.

2.1 Les Microblogs

Dans le chapitre 6, nous montrons quel potentiel constituent les microblogs pour l'analyse de sentiments. Une mode récemment apparue sur Internet a engendré une explosion du nombre de sites permettant de diffuser des microblogs, ces petits messages dont la taille maximale est restreinte à un texte très court. A tel point que c'est devenu en quelques années un des principaux type de communication sur Internet. La quantité très importante d'information présente sur les sites de microblogs les rend attractifs en tant que source de données pour la fouille d'opinion et l'analyse de sentiments. Nous utilisons Twitter, la plus grande plateforme de microblogs à ce jour, comme une source de données multilingues pour l'analyse de sentiments. Dans le chapitre 6, nous montrons comment nous avons obtenu un jeu de données étiqueté avec des annotations décrivant les sentiments exprimés dans les blogs, de manière automatique, en utilisant les émoticônes[2] comme des annotations bruitées. Nous relatons ensuite comment nous avons utilisé ces données dans 3 types de tâches :

1. construction d'un lexique affectif pour différentes langues,
2. classement en polarité de critiques de jeux vidéo en français,
3. désambiguïsation d'adjectifs ambigus exprimant des sentiments en chinois.

Notre intention était de ne pas utiliser d'outil linguistique sophistiqué, afin de préserver à notre approche son caractère indépendant de la langue cible, ou du moins facilement transposable à une autre langue. Nous avons évalué notre approche en effectuant des expériences par comparaison des résultats avec ceux obtenus en utilisant un corpus annoté manuellement, ou un lexique construit par des experts ainsi qu'en participant à la campagne d'évaluation internationale SemEval 2010[3]. Lors de ces évaluations notre système a obtenu des performances comparables à celles d'un classifieur supervisé, qui lui nécessite de disposer de données d'apprentissage annotées. Notre méthode est entièrement automatique et n'a besoin d'aucune autre ressource langagière construite à la main.

2.2 Les d-grammes

Les modèles n-grammes sont un moyen traditionnel de représentation des textes, souvent utilisé en analyse de sentiments. Cependant, nous pensons que la difficulté intrinsèque de la tâche appelle à l'utilisation de nouveaux modèles mieux adaptés à la capture des opinions. C'est pourquoi nous proposons dans le chapitre 7 un nouveau modèle s'inspirant des n-grammes, mais construits à partir des triplets constitutifs des dépendances syntaxiques. Nous avons appelé ce nouveau modèle : d-gramme. De nos expériences, il ressort que l'approche à base de d-grammes contient plus d'information pertinente pour l'analyse de

[2]Ces esquisses de visages représentant une émotion ou un états d'esprit particulier au moyen de quelques caractères qui sont fréquemment utilisés dans les communications sur Internet.

[3]Cette participation est présentée en détails au chapitre 9).

sentiments que les simples modèles à sac-de-mots. Prenons comme exemple l'énoncé « La bande son était affreuse » :

$$S = \text{“The soundtrack was awful”}$$

Un graphe de dépendances syntaxiques possible pour cet énoncé est présenté dans la figure 1.

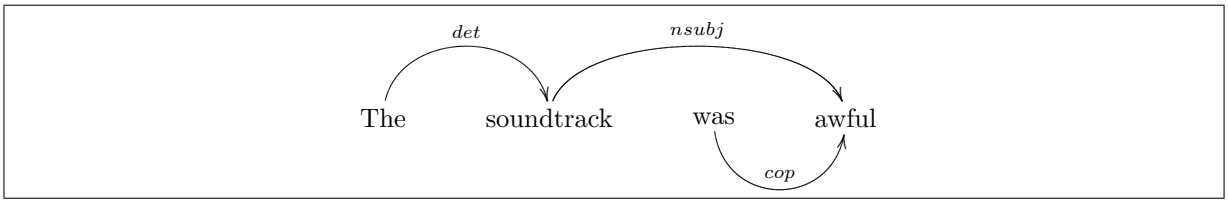

Figure 1: *Le graphe de dépendances syntaxiques produit par l'analyseur syntaxique de Stanford pour l'énoncé “The soundtrack was awful”*

À partir des dépendances de ce graphe nous construisons les d-grammes suivants :

$$\begin{aligned}\text{dgrams}(S) = \{&(\text{The}, \text{det}, \text{soundtrack}),\\ &(\text{soundtrack}, \text{nsubj}, \text{awful}),\\ &(\text{was}, \text{cop}, \text{awful})\}\end{aligned}$$

À fin de comparaison, voici les représentations à base respectivement d'unigrammes et de bigrammes pour le même énoncé :

$$\begin{aligned}\text{unigrams}(S) = \{&(\text{The}),\\ &(\text{soundtrack}),\\ &(\text{was}),\\ &(\text{awful})\}\end{aligned}$$

$$\begin{aligned}\text{bigrams}(S) = \{&(\text{The}, \text{soundtrack}),\\ &(\text{soundtrack}, \text{was}),\\ &(\text{was}, \text{awful})\}\end{aligned}$$

Les modèles d-grammes présentent l'avantage d'être capable de trouver les dépendances à longue distance et d'apporter une information plus pertinente pour le rattachement syntaxique des mots entre eux. Fait important pour la fouille d'opinion, les modèles d-grammes facilitent aussi le repérage des négations.

Pour nos évaluations nous avons utilisé le jeu de données Cross-Lingual Sentiment[4] et construit par [8]. Ce jeu de données est composé de critiques de produits commerciaux publiées sur Amazon dans 4 langues. Nous avons utilisé les critiques en anglais et en français. Elles sont réparties en 3 domaines selon le type de produit : livres, musique et DVDs. Pour chaque domaine nous avons 2.000 critiques positives et 2.000 négatives, pour un total de 24.000 documents utilisés dans cette expérience. Pour évaluer notre modèle nous avons utilisé une implémentation de machine à vecteurs supports linéaire de la bibliothèque LIBLINEAR [4] et une implémentation personnelle d'un classifieur Bayésien naïf . Nous avons effectué un validation croisée à 10 replis pour estimer l'exactitude moyenne (average accuracy). Les résultats de l'expérience montrent l'efficacité de l'approche d-gramme par rapport aux modèles traditionnels à base d'unigrammes ou de bigrammes. Nous en concluons donc que notre méthode est générale et indépendante du domaine d'application ou de la langue cible.

2.3 Améliorer les stratégies de pondération

Le chapitre 8 présente le problème de l'écrasement des statistiques des opinions minoritaires par celles des opinions majoritaires dans les approches classiques et la solution que nous proposons pour résoudre

[4]Cross-Lingual Sentiment disponible à l'url :
`http://www.uni-weimar.de/cms/medien/webis/research/corpora/webis-cls-10.html`

ce problème. De nos jours, il est très facile de rassembler de très grosses quantités de textes contenant des opinions à partir d'Internet, en allant les chercher sur les réseaux sociaux, les sites de critiques de produits commerciaux, les forums etc. Il est possible à partir de ces corpus de construire assez facilement un système de classement en polarité, qui soit capable de classer des documents de même nature que ceux du corpus original, avec un niveau acceptable d'exactitude. Cependant le système ainsi obtenu comporte un biais en faveur des opinions majoritairement exprimées dans le corpus d'apprentissage. Si maintenant, nous utilisions ce système pour déterminer la polarité de nouvelles critiques d'un produit pour lequel les critiques positives sont majoritaires dans le corpus d'apprentissage, il est fortement probable que toutes les nouvelles critiques ainsi analysées seraient aussi considérées comme positives, car le texte de ces critiques contiendra certainement les mêmes traits considérés comme indicateurs de positivité (par exemple le nom du produit, la marque ou la référence au modèle du produit) que les critiques du même produit dans le corpus d'apprentissage. Et ce, quel que soit le domaine applicatif considéré, par exemple pour les films, ces traits indicateurs de positivité seraient le titre, le nom des acteurs principaux, le nom du réalisateur, du producteur etc. Si nous appelons la cible de l'opinion l'*entité* (entity), ces traits peuvent être considérés comme étant *spécifiques à l'entité* (entity-specific). Paradoxalement, ce biais est un facteur d'amélioration de l'exactitude du système de classement en polarité, car la distribution entre les critiques positives et négatives d'un produit est en général la même entre le corpus d'apprentissage et le corpus de test sur lequel on applique l'algorithme de classement. Si un produit est bon, il recevra aussi plus de critiques positives dans ce nouveau corpus et réciproquement.

Cependant, nous pouvons être amené à souhaiter disposer d'un système de classement en polarité, qui non seulement possède globalement (en moyenne) une bonne exactitude, mais qui soit aussi capable de déterminer correctement la polarité d'une critique minoritaire, c'est à dire des critique, qui malgrè les louanges de la majorité notent très négativement un produit ou au contraire mentionnent des points positifs pour un produit considéré comme mauvais. Le fonctionnement d'un tel système, s'approche plus de celui d'un expert, qui prend une décision objective à partir des informations mentionnées dans la le texte de la critique, sans se laisser influencer a priori par les opinions que la majorité entretient pour ce produit.

Pour prouver nos dires, nous avons utilisé deux jeux de données standards: des critiques de film et de produits commerciaux qui ont été utilisé par le passé pour des recherche en analyse de sentiment [5], [1], [3]. Pour chacun des deux jeux de donnée, nous avons construit une version *biaisée* (biased), en regroupant d'abord les critiques en fonction de leur entité cible (un film ou un produit particulier) et ensuite en sélectionnant des groupes avec une distribution déséquilibrée entre les critiques positives et négatives. Nous montrons que les approches traditionnelles, c'est-à-dire utilisant des machines à vecteurs supports et des n-grammes de traits, sont beaucoup moins performantes pour classer les critiques minoritaires que pour classer les critiques majoritaires dans notre jeu de données biaisé. Pour améliorer le classement des critiques minoritaires, nous devons réduire l'importance des termes qui pourraient introduire un biais dans le classement, ce que nous proposons de faire au moyen de deux mesures : la *fréquence moyenne d'un terme* et la *proportion d'entité*.

2.3.1 Fréquence moyenne d'un terme

La fréquence moyenne d'un terme (agv.tf) est le nombre moyen de fois qu'un terme apparaît dans un document :

$$\text{avg.tf}(g_i) = \frac{\sum_{\{d|g_i \in d\}} \text{tf}(\text{g}_\text{i})}{\|\{d|g_i \in d\}\|} \tag{1}$$

où $\{d|g_i \in d\}$ est l'ensemble des documents qui contient g_i

La normalisation à base de fréquence moyenne d'un terme est basée sur l'observation que les auteurs de critiques ont tendance à utiliser un vocabulaire riche quand ils expriment leur attitude par rapport à un film ou un produit. Ainsi, les termes expriment des sentiments comme *remarquable* (outstanding) ou *adorable* (lovingly) ont une fréquence moyenne proche ou égale à 1, tandis que les termes non subjectifs ont une fréquence moyenne plus élevée. En particulier, cela est vrai pour les termes spécifiques à l'entité comme les titres de film, les noms d'acteurs, les marques et les noms de modèles qui sont souvent

mentionnés plusieurs fois au sein d'un même document. Afin de normaliser le vecteur représentatif d'un document qui associe à chaque terme présent dans le document un poids représentatif de son importance, nous divisons chaque poids par la fréquence moyenne du terme correspondante (avg.tf) :

$$w(g_i)^* = \frac{w(g_i)}{\text{avg.tf}(g_i)} \tag{2}$$

2.3.2 Proportion d'entité

La proportion d'entité (ep) est la proportion des occurrences d'un terme par rapport aux différentes entités comparativement à la fréquence des documents :

$$\text{ep}(g_i) = \log\left(\frac{\|\{e|g_i \in e\}\|}{\|\{d|g_i \in d\}\|} \cdot \frac{\|D\|}{\|E\|}\right) \tag{3}$$

où $\{e|g_i \in e\}$ est l'ensemble des entités qui contient g_i dans leurs critiques, $\|D\|$ est le nombre total de documents, $\|E\|$ est le nombre total d'entités.

La normalisation de proportion d'entité favorise les termes qui apparaissent dans les critiques de nombreuses entités mais dans peu de documents. Nous distinguons trois types de termes :

1. Les termes spécifiques à une entité, tels que les noms de produits ou les titres de film, qui sont associés à peu d'entités et donc devraient apparaître dans peu de documents. La valeur de ep devrait être proche de celle de la constante de normalisation $\frac{\|D\|}{\|E\|}$ (nombre moyen de documents par produit).
2. Les termes subjectifs, tels que « remarquable » (“outstanding”) ou « adorable » (“lovingly”), qui devraient apparaître associés à beaucoup de produits et dans un nombre relativement restreint de documents, car les auteurs utilisent un vocabulaire varié. La valeur de ep sera plus grande que la constante de normalisation.
3. Les mots-outils, tels que les déterminants et les prépositions, devraient apparaître dans presque tous les documents, et donc associés à presque tous les produits. La valeur de ep sera proche de celle de la constante de normalisation.

Pour normaliser le vecteur représentatif d'un document, nous multiplions chaque poids associé à un terme par la proportion d'entité associée à l'objet de la critique.

$$w(g_i)^* = w(g_i) \cdot \text{ep}(g_i) \tag{4}$$

Toutes nos expériences réalisées avec des versions spécialement préparées de jeux de données standards ont montré une amélioration des performances de l'exactitude de classification pour les critiques minoritaires. Cependant nous avons quand même observé une légère baisse dans la mesure d'exactitude globale, car il est toujours plus bénéfique pour un classement en polarité de suivre la tendance majoritaire d'opinion.

Au final, c'est toujours le développeur d'un système de classement en polarité qui devra choisir entre un système biaisé dont la performance globale sera légèrement meilleure et un système capable d'identifier correctement les textes d'opinion minoritaire en fonction des visées applicatives qui président à la création de son système. Des applications possibles de notre procédure de normalisation sont l'analyse des retours clients, afin de détecter très tôt les critiques pour supprimer la source du mécontentement ou la détection des signaux faibles (rumeurs), en particulier pour les application sécuritaires, des types d'application qui ont besoin d'une classification à grain fin des documents.

3 Applications

Un des aspects important de nos travaux concerne leur application directe à des problèmes concrets. Dans la partie 3 de notre thèse, nous relatons notre participation à différente campagnes d'évaluation,

qui nous a permis de tester notre approche. Plus particulièrement, nous avons participé aux campagnes internationales suivantes :

- SemEval'10 : désambiguïsation d'adjectifs ambigus exprimant des sentiments en chinois (chapitre 9)
- ROMIP'11 : classement en polarité de critiques de produits commerciaux en russe (chapitre 10)
- I2B2'11 : détection d'émotions dans des notes de suicide (chapitre 11)

3.1 SemEval'10

Le jeu de données de la campagne SemEval 2010 [9] est constitué de courts textes en chinois contenant des adjectifs pris dans une liste fermée, et dont le sentiment associé doit être désambiguïsé en fonction du contexte. Dans notre approche, nous utilisons la plate-forme de microblogs Twitter pour collecter des messages à teneur émotionnelle et construire deux sous-corpus contenant respectivement : les messages à teneur positive et ceux à teneur négative, comme nous l'avons décrit dans le chapitre 6. Notre système de classement en sentiments construit avec ces données, utilise une approche bayésienne naïve multinomiale. Puisque les textes à analyser étaient courts, nous fait l'hypothèse que la polarité à associer à l'adjectif était la même polarité que celle du document entier.

Dans nos expériences, nous avons utilisé deux jeux de données : un jeu d'essai contenant 100 phrase en chinois et un jeu de test de 2.917 phrases, tous deux fournis par les organisateurs de la campagne d'évaluation. Les mesures d'évaluation utilisées pour cette campagne sont la micro et la macro exactitude. Il faut noter que notre approche peut être appliquée sans changement à n'importe quelle autre langue à condition de disposer de suffisamment de données d'apprentissage issues de Twitter. Nous avons obtenu respectivement 64% de macro et 61% de micro exactitude, à la tâche de SemEval 2010, ce qui est une performance inférieure à celle de la plupart des autres participants (nous sommes $6^{\text{èmes}}$ sur 7 participants), mais notre système est entièrement automatique et n'a recourt à aucun lexique construite manuellement.

3.2 ROMIP'11

ROMIP est une campagne internationale d'évaluation annuelle en recherche d'information qui a débuté en 2002 [2]. Pour la campagne de 2011, les organisateurs ont ajouté une piste sur l'analyse de sentiments dont le but était le classement en opinion de textes écrits par des consommateurs. Un jeu de données composé de critiques de produits commerciaux issues du services en ligne de recommandation Imhonet et de l'aggrégateur de produits Yandex.Market a été fourni aux participant pour entraîner leurs systèmes. Le jeu de données contenait des critiques pour trois types de produits : les appareils photo numériques, les livres et les films.

L'analyse de sentiment est une tâche difficile, même pour les langues pour lesquelles les ressources linguistiques sont nombreuses comme c'est le cas pour l'anglais. En plus des traitements relativement basiques comme l'étiquetage morpho-syntaxique, des outils d'analyse du langage plus sophistiqués comme des analyseurs de discours ou des lexiques spécifiques sont nécessaires à certaines approches actuelles. C'est pourquoi il est très difficile d'adapter des méthodes initialement développées pour d'autres langues au russe, en particulier celles développées pour l'anglais. L'un des rares système de classement en sentiment développé pour la langue russe [7] est un système à base de règles, qui utilise un lexique affectif construit manuellement ainsi qu'un étiquetage morpho-syntaxique et des informations syntaxiques au niveau lexical. Cependant, à notre connaissance, il n'existe pas de ressource publique pour l'analyse de sentiments en russe et développer une approche à base de lexique serait bien trop coûteuse car il faudrait partir de rien.

Pour résoudre ce problème, nous avons décidé d'employer une approche indépendante de la langue qui n'ait pas besoin d'analyse du traitement du langage sophistiquée ni de lexique dédié, qui rappelons le, n'existe pas pour le russe. C'est pourquoi, nous avons employé un système à base de machine à vecteurs supports avec des traits construits sur des n-grammes, des étiquettes morpho-syntaxiques et une

analyse syntaxique en dépendances. Nous avons du entraîner un analyseur syntaxique en dépendances sur le corpus national russe (Russian National Corpus) [5]. De plus nous avons effectué une étude sur la pondération des termes et la composition du corpus pour optimiser les performances de notre système qui a été classé d'après les mesures de performance officielles, quatrième dans la piste de classification binaire pour le domaine des appareils photo numériques, troisième dans la piste à trois classes pour le domaine des films et premier dans la piste à cinq classes tous domaines confondus,

3.3 I2B2'11

La seconde piste de la campagne d'évaluation I2B2 2011 avait pour but la reconnaissance des opinions exprimées dans un corpus de notes de suicide, en étiquetant les phrases à l'aide d'une ou plusieurs des quinze catégories suivantes : *instructions* (instructions), *information*(information), *désespoir* (hopelessness), *culpabilité* (guilt), *reproche* (blame), *colère* (anger), *chagrin* (sorrow), *peur* (fear), *maltraitance* (abuse), *amour* (love), *reconnaissance* (thankfulness), *espoir* (hopefulness), *bonheur-tranquilité* (happiness-peacefulness), *fierté* (pride), *pardon* (forgiveness).

Nous avons contribué au développement d'un système combinant des règles manuelles et une approche d'apprentissage automatique pour détecter les émotions. Notre objectif était de créer un système qui possède la précision des systèmes à base de règles, secondé par des algorithmes d'apprentissage automatique pour améliorer le rappel et les capacités de généralisation à de nouvelles données. Notre contribution a concerné l'apprentissage automatique, pour lequel nous avons entraîné un système de classement à vecteurs supports utilisant différents traits extraits des corpus d'apprentissage. Nous avons utilisé la bibliothèque LIBLINEAR [4] avec un noyau linéaire et un paramétrage par défaut. Pour la classification multi-étiquettes, nous avons utilisé une stratégie en parallèle, c'est-à-dire que nous avons entraîné indépendamment un système de classement pour chaque émotion. Chaque système de classement fournit pour chaque phrase une indication de présence ou d'absence de l'émotion qu'il a été entraîné à détecter. Ainsi nous pouvons obtenir pour chaque énoncé, de 0 à 15 étiquettes d'émotion. La liste des traits utilisés pour l'apprentissage comprenait : des n-grammes, des d-grammes (chapitre 7), des étiquettes morpho-syntaxiques, des traits de la base *General Inquirer* (décrite dans le chapitre 5), des traits de la base ANEW (décrite dans le chapitre 5) et des traits heuristiques complémentaires (par ex. la position de la phrase dans la note).

Au final, notre algorithme de classement était le suivant :

1. D'abord nous avons entraîné un détecteur d'annotation, pour distinguer les phrases qu'il fallait annoter des phrases qui resteraient dépourvues d'annotation. Les traits utilisés ont été : les étiquettes morpho-syntaxiques et les traits *General Inquirer*.

2. Ensuite, les phrases qui étaient supposées recevoir des annotations étaient traitées par un détecteur de subjectivité, afin de séparer les phrases objectives de celles subjectives. Les traits utilisés ont été : les traits heuristiques complémentaires, les étiquettes morpho-syntaxiques et les traits *General Inquirer*.

3. Parmi les phrases objectives, nous avons identifié celles qui contenaient des *information* et celles qui contenaient des *instructions*. Les traits utilisé ont été : les unigrammes, les bigrammes, les traits *General Inquirer*, les graphes de dépendances syntaxiques.

4. Les phrases subjectives ont été réparties en deux classes, celles qui contenaient des émotions positives et celles qui contenaient des émotions négatives. Les traits utilisés ont été : les étiquettes morpho-syntaxiques et les traits ANEW.

5. Les phrases à connotation négative ont ensuite été réparties entre les 7 classes suivantes: *chagrin* (sorrow), *désespoir* (hopelessness), *maltraitance* (abuse), *culpabilité* (guilt), *reproche* (blame), *peur* (fear), *colère* (anger). Les traits utilisé ont été: les unigrammes, les bigrammes, les d-grammes, et les traits *General Inquirer*.

[5] **Russian National Corpus**: `http://www.ruscorpora.ru/en/`

6. Les phrases avec une polarité positive ont été elles réparties entre les 6 classes suivantes: *fierté* (pride), *espoir* (hopefulness), *amour* (love), *bonheur-tranquilité* (happiness-peacefulness), *reconnaissance* (thankfulness), *pardon* (forgiveness). Les traits utilisés ont été : les unigrammes, les bigrammes, les d-grammes et les traits *General Inquirer*.

Afin d'affiner le paramétrage de la partie apprentissage automatique de notre système de classement, nous avons effectué une validation croisée à 10 replis sur le corpus d'apprentissage. Les mesures de performance officielles étaient: la micro-moyenne en précision, en rappel , et en F-mesure (score F1). La moyenne officielle des F-score était 0.4875, le plus mauvais F-score 0.2967 et le meilleur 0.6139. L'approche à base de règles seule a obtenu : F-score = 0.4545, précision = 0.5662, rappel = 0.3797, tandis que notre meilleur paramétrage de l'approche combinant règle et apprentissage automatique a obtenu : F-score = 0.5383, précision = 0.5381 et rappel = 0.5385. Notre approche combinée a été classée sixième sur vingt-six.

4 Conclusion

La reconnaissance des sentiments dans les textes oblige les chercheurs à se confronter à de nombreuse questions d'analyse du langage, telles que l'analyse du discours, la résolution des coréférences, la reconnaissance des métaphores etc. Dans nos travaux, nous nous sommes intéressés au classement en polarité, une des tâches fondamentales de l'analyse de sentiments qui vise à classer les documents en fonction de l'attitude qu'a le détenteur d'une opinion envers la cible de l'opinion. Même dans un cadre simplifié, cela reste une tâche difficile, d'autant plus que l'on fonctionne dans un environnement multilingue ou multi-domaines. Notre approche cherche à créer un système d'analyse de sentiments automatique et adaptatif qui soit indépendant de la langue et du domaine d'application concerné.

Notre contribution porte sur les éléments suivants :

- Nous avons montré comment utiliser les microblogs comme une source de données multilingue pour la fouille d'opinion et l'analyse de sentiments. Dans nos expériences, nous nous sommes servi de Twitter pour collecter un corpus de messages exprimant des opinions.
- Nous avons proposé une méthode automatique pour étiqueter les messages de Twitter comme positif ou négatif en considérant les émoticônes présentes dans les messages, comme des étiquettes bruitées. Nous avons ainsi obtenu un ensemble de messages positifs et négatifs pour quatre langues : 5,2 millions de messages en anglais, 672.800 en espagnol, 287.400 en français et 7.800 en chinois. Pour l'anglais, nous avons collecté un ensemble additionnel de 100.000 messages considérés comme neutres en polarité à partir des messages publiés sur Tiwtter par les journaux.
- Nous avons effectué une analyse linguistique du corpus collecté et observé que la distribution des étiquettes morpho-syntaxiques est différente entre le sous-corpus de messages subjectifs et le sous-corpus de messages objectifs, ainsi qu'entre le sous-corpus de messages de polarité positive et celui contenant les messages de polarité négative. Nous proposons d'utiliser la distribution des étiquettes morpho-syntaxiques comme trait supplémentaire pour le classement en polarité et la détection de subjectivité.
- Nous avons proposé une méthode pour la construction automatique de lexiques affectifs à partir de Twitter. Nous avons ainsi construit des lexiques pour l'anglais, l'espagnol et le français qui ont été évalués en vérifiant la corrélation des informations qu'ils contenaient avec le contenu de la base ANEW, considérée comme un standard du domaine.
- Nous avons utilisé les lexiques produits précédemment pour le classement en polarité de critiques de jeux vidéo en français dans le cadre des évaluations effectuées pour le projet DOXA. Les résultats de l'évaluation ont montré que les performances de notre approche sont comparables à celles obtenues avec une approche à base d'apprentissage automatique supervisé utilisant des n-grammes, alors que notre approche n'a pas besoin de corpus d'apprentissage, car elle se satisfait des ressources extraites automatiquement à partir de Twitter.

- Nous avons proposé un nouveau mode de représentation des textes pour l'analyse de sentiments, que nous avons baptisé d-grammes et qui est basé sur les graphes de dépendances syntaxiques. Des évaluations effectuées avec trois analyseurs syntaxiques en dépendances différents, sur un jeu de données multi-domaines de critiques de produits en anglais et en français, ont montré que notre modèle de d-grammes permet d'obtenir une meilleure exactitude en classement de polarité; avec une amélioration de score pouvant aller jusqu'à 4,4% par rapport à l'approche traditionnelle à base de n-grammes.

- Nous avons exhibé une faiblesse de l'approche traditionnelle pour le classement supervisé en polarité pour ce qui concerne le classement des opinions minoritaires. Nous avons montré que les systèmes de classement ont tendance à s'appuyer sur les traits spécifiques aux entités cibles des opinions et, par voie de conséquences, qu'ils sont biaisés en faveur des opinions majoritaires.

- Nous avons proposé deux mesures pour normaliser les poids qualifiant l'importance d'un terme pour un classement en opinion : la fréquence moyenne d'un terme et la proportion d'entité. Des évaluations effectuées sur deux jeux de données en anglais, concernant cinq domaines applicatifs (films, livres, DVDs, électroménager, électronique) ont montré une amélioration des performance de classification des opinions minoritaires pouvant aller jusqu'à 12,5%.

Les approches proposées pour l'analyse de sentiment automatique et adaptative ont été testées avec succès dans les campagnes d'évaluation internationales suivantes :

- ROMIP 2011 : concernait le classement en polarité sur des critiques de produits commerciaux en russe. Parmi vingts systèmes participants, notre système a été classé quatrième dans la tâche de classification binaire pour le domaine des appareil photo électroniques, troisième dans la piste à trois classes pour le domaine des films, et premier dans la piste à cinq classes tous domaines confondus.

- SemEval 2010 : qui portait sur la désambiguïsation d'adjectifs ambigus exprimant des sentiments en chinois. Avec notre approche indépendante du langage, nous avons obtenu 64% de macro et 61% de micro exactitude (accuracy) et avons été classé sixième sur sept participants.

- I2B2 2011 : qui traitait de la détection des émotions dans des notes de suicide. Notre système a été classé sixième sur 26 avec une F-mesure de 0.5383, qui était bien supérieure à la moyenne officielles des scores obtenus qui était de 0.4875.

Nous pensons que notre approche peut être facilement combinée avec d'autres travaux du domaine de la fouille d'opinion et l'analyse de sentiments parce que nos algorithmes sont facilement portables vers de nouveaux domaines applicatifs ou de nouvelles langues.

Bien que nous nous soyons concentrés dans cette thèse uniquement sur la classification de polarité, pour nos travaux futurs, nous envisageons le développement d'une approche combinée pour le problème de la fouille d'opinion et l'analyse de sentiments qui tienne compte de manière conjointe de tous les paramètres informationnels d'une expression d'opinion, c'est à dire l'expression, sa cible et sa source explicite ou implicite.

Les travaux présentés dans cette thèse ont fait l'objet à ce jour de publications dans 2 revues internationales, 5 conférences, 4 ateliers internationaux, et 1 chapitre de livre.

[1] J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[2] B. Dobrov, I. Kuralenok, N. Loukachevitch, I. Nekrestyanov, and I. Segalovich. Russian Information Retrieval Evaluation Seminar. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal, May 2004.

[3] K. Duh, A. Fujino, and M. Nagata. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 429–433, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[4] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[5] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[6] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, January 2008.

[7] A. G. Pazelskaya and A. N. Solovyev. A method of sentiment analysis in Russian texts. In *Proceedings of the Dialog 2011 the 17th International Conference On Computational Linguistics*, Moscow region, Russia, May 2011.

[8] P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1118–1127, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[9] Y. Wu, P. Jin, M. Wen, and S. Yu. Semeval-2010 task 18: Disambiguating sentiment ambiguous adjectives. In *SemEval 2010: Proceedings of International Workshop of Semantic Evaluations*, pages 275–278, Uppsala, Sweden, 2010. ACL SigLex.