Exploitation de lexiques pour la catégorisation fine d'émotions, de sentiments et d'opinions

Nicolas Hernandez, Grégoire Jadi, Joseph Lark, Laura Monceaux

Exploitation de lexiques pour la catégorisation fine d'émotions, de sentiments et d'opinions

Nicolas Hernandez[1] Grégoire Jadi[1] Joseph Lark[1,2] Laura Monceaux[1]
(1) LINA, UMR CNRS 4261, 2 chemin de la Houssinière, Nantes, France
(2) Dictanova, 2 chemin de la Houssinière, Nantes, France
{Prénom.Nom}@univ-nantes.fr

Résumé. Nous présentons dans cet article notre proposition pour la 11ème édition du Défi Fouille de Textes (DEFT). Nous participons à trois tâches proposées dans le cadre de cet atelier en fouille d'opinion. Les objectifs de ces tâches sont de classer des tweets en français sur le sujet des énergies renouvelables, respectivement du point de vue de la polarité, du type général d'information énoncé, et enfin de la classe fine du sentiment, de l'émotion ou de l'opinion exprimée. Pour réaliser cette catégorisation, nous proposons d'explorer et d'évaluer différentes méthodes de construction de lexiques typés sémantiquement : outre des lexiques affectifs construits manuellement, nous expérimentons des lexiques typés construits semi-automatiquement sur le corpus d'évaluation et d'autres sur un corpus tiers.

Abstract.
Using affective lexicons for fine-grained sentiment, opinion and emotion analysis
In this article, we present our contribution to the 11th DEFT workshop (Défi Fouille de Textes). We take part in three tasks proposed in this opinion mining challenge. The goal of these tasks is to analyse a corpus of french tweets about renewable energy, through the inference of their polarity, general semantic class, and fine-grained sentiment class respectively. Our proposition makes use of a machine learning process that combines various ways of building semantically classified lexicons. We explore the use of external lexicons, semi-supervisely extracted lexicons from the training corpus, and semi-supervisely extracted lexicons from a third-party corpus.

Mots-clés : Fouille d'opinion, expression d'émotions, analyse de sentiments, construction de lexique, classification fine.

Keywords: Sentiment analysis, opinion mining, fine-grained emotion classification, lexicon acquisition.

1 Introduction

Avec l'expansion du web social, les internautes sont de plus en plus enclins à partager leurs avis sur les réseaux sociaux ou les sites spécialisés. Le domaine de la fouille d'opinion vise à désambiguïser automatiquement ces informations en ce qui concerne le sentiment exprimé, en le traduisant par une valence affective (polarité "positive" ou "négative", score de subjectivité...) ou par une catégorie de sentiment. C'est dans cette optique que notre travail s'inscrit, répondant ainsi à la tâche de classification fine de tweets (tâche 2.2) proposée pour la compétition DEFT 2015. Nous avons inféré les informations plus générales (tâches 1 et 2.1) depuis nos résultats en catégorisation fine.

La classification demandée compte 18 catégories sémantiques, à laquelle s'ajoute une classe "neutre" correspondant à l'énoncé d'une information objective. Ces catégories détaillées impliquent des variations relativement sensibles entre les classes pouvant entraîner des ambiguïtés, ce qui constitue la difficulté majeure de cette tâche. En particulier, la résolution de ces ambiguïtés peut être complexe dans le cas où deux catégories sémantiquement proches ne sont pas représentées de façon équilibrée, car il peut exister un biais en faveur de la classe la plus présente. Afin de catégoriser les tweets, nous avons, tout d'abord, utilisé une représentation en sac de bigrammes de mots. Nous avons par la suite amélioré ces premiers résultats au moyen (1) de lexiques construits manuellement, (2) de l'acquisition semi-automatique de mots sémantiquement liés aux catégories définies au sein d'un corpus externe ou (3) au sein du corpus d'entraînement.

Dans la suite de cet article, nous dressons un bref état de l'art des méthodes liées aux problématiques soulevées par ce défi (section 2), puis nous exposons notre démarche (section 3) ainsi que les détails des méthodes utilisées (section 4). Ensuite, nos présentons les résultats observés sur le corpus fourni (section 5). Enfin, nous commentons ces résultats, et le travail effectué lors de cette participation en général (section 6).

2 Travaux connexes

Ce travail est selon nous fortement lié à la détection de subjectivité : il s'agit en effet de déterminer dans quelle mesure l'information présente dans un tweet renvoie à un sentiment ou une émotion. Nous considérons dans ce contexte qu'une première différenciation peut être faite entre la classe neutre du point de vue de la subjectivité ("Information") et toutes les autres. Ce type de différenciation fait l'objet de plusieurs travaux dans la littérature (Wiebe & Mihalcea, 2006; Murray & Carenini, 2011). Cependant la notion de subjectivité peut recouvrir plusieurs concepts. Liu (2012) distingue ainsi les expressions d'un désir, d'une opinion, d'une croyance, d'une spéculation... et montre que l'on peut être amené à confondre la subjectivité d'une phrase et le fait qu'elle exprime un sentiment ou une opinion. C'est sur ce plan que l'on peut distinguer par exemple un jugement rationnel d'une opinion passionnée. À ces modalités de l'expression d'un sentiment s'ajoutent les différentes émotions humaines. Parrott (2001) identifie six émotions primaires que sont la joie, l'amour, la surprise, la colère, la tristesse et la peur. L'ensemble de ces catégories d'expression sont représentées par les classes que nous cherchons à identifier ici. Les travaux réalisant une identification similaire reposent pour la plupart sur des lexiques affectifs. Dans cette optique, Staiano & Guerini (2014) exploitent un corpus journalistique annoté par les internautes selon l'émotion suscitée par chaque nouvelle afin d'en extraire un lexique d'émotion de près de 37 000 mots. Yang *et al.* (2014) utilisent une version de l'allocation de Dirichlet latente (LDA) pour rechercher des mots sémantiquement proches de graines définies manuellement exprimant une émotion. En français, Vernier *et al.* (2009) proposent une méthode d'apprentissage automatique des structures caractéristiques d'une évaluation dans le texte.

3 Approche globale

Nous décrivons ici notre position sur les différentes tâches et notre approche pour intégrer les différents traits observés.

3.1 Appréhension des différentes tâches

Nous avons considéré la tâche 2.2 (identification de la classe spécifique de l'opinion, sentiment ou émotion d'un tweet donné) comme une spécialisation de la tâche 2.1 (identification de la classe générique de l'information exprimée dans le tweet), et celle-ci comme une spécialisation de la tâche 1 (classification des tweets selon leur polarité). En consacrant nos efforts sur la tâche 2.2, nous avions ainsi par différents degrés de généralisation la possibilité d'obtenir des résultats pour les tâches moins spécifiques.

3.2 Intégration des différents traits par apprentissage

Comme indiqué dans (Pustejovsky & Stubbs, 2012), les modèles génératifs de type bayésien naïf (NB) ou discriminatifs de type machines à vecteurs de support (SVM) ou d'entropie maximale sont connus pour être mieux adaptés que d'autres sur des tâches de classification d'énoncés selon un jeu de catégories.

La disponibilité d'un corpus d'entrainement, la taille des données de test et la multiplicité des classes à reconnaître nous a conduit vers la voie de l'apprentissage supervisé. Nous n'étions néanmoins pas réfractaires à la mise en place de post-traitements à base de règles pour corriger des biais identifiés.

Du fait de l'hétérogénité des données (déséquilibre des classes, faible représentativité de certaines), et au regard du nombre d'instances et de traits (de quelques centaines à plusieurs milliers) que nous souhaitions considérer (possiblement quelques milliers chacun), nous avons opté pour l'utilisation d'un classifieur linéaire de type SVM comme conseillé dans la littérature (Hsu *et al.*, 2003). Ce type de classifieur offre la possibilité de manipuler plusieurs milliers de traits et d'obtenir des modèles parmi les plus performants en seulement quelques secondes [1]. Nous avons utilisé en particulier l'implémentation offerte par (Yu *et al.*, 2013) qui exploite sur la bibliothèque `LIBLINEAR` [2] (Fan *et al.*, 2008). Nous avons opté pour une représentation binaire des traits, une normalisation des instances et un classifieur multi-classes (Crammer & Singer, 2000). Ce paramétrage correspond au mode par défaut ; celui-ci produisait les meilleurs résultats sur nos données. Alors qu'une représentation des traits en nombre d'occurrences était légèrement moins bon la représentation en fonction de la

1. En comparaison, un NB prend quelques minutes et un arbre de décision type C4.5 plusieurs heures avec un corpus comptant approximativemet 6 000 instances décrites chacunes par environ 80 000 traits.

2. `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`

fréquence des termes ou de la fréquence des termes plus l'inverse de la fréquence documentaire, donnaient de moins bons résultats sur les données d'entraînement. Le peu de matériel lexical des énoncés à annoter et la taille modeste du corpus (voire très petite pour certaines classes) expliquent l'inadéquation de ces représentations à notre tâche.

Pour estimer les paramètres optimaux (notamment le coût de la pénalité C), nous avons procédé par une recherche "par quadrillage" qui consiste à tester différentes valeurs (incrémentées exponentiellement) et à sélectionner celles qui donnent les meilleurs résultats par validation croisée sur cinq strates. Un autre argument pour motiver notre choix pour un SVM venait du fait que n'ayant aucune information sur la distribution des classes dans le corpus de test, nous souhaitions un classifieur robuste à une distribution différente. Certains classifieurs comme le NB prend en compte les probabilités d'occurrence observées dans le corpus d'entraînement. Nous craignions qu'avec un tel classifieur les classes prédites soient seulement les dominantes de notre corpus d'entraînement.

4 Descriptions des différents traits considérés

Dans cette section, nous décrivons notre approche de base et les différentes approches reposant sur des lexiques existants, des lexiques construits sur le corpus d'entraînement (endogènes)et d'autres sur un corpus tiers (exogènes).

4.1 Un modèle de bigrammes de tokens mots

Nous choisissons une modélisation en bigrammes de tokens mots avec une représentation binaire comme approche de base. Sur le corpus d'entraînement, cela correspond à 86 018 traits distincts pour 21 295 tokens uniques. Ces bigrammes de tokens mots ont été calculés sans normalisation du texte (surface brute, aucune tokenization ni racinisation).

4.2 Construction et utilisation de lexiques exogènes

Nous avons projeté au corpus d'entraînement trois lexiques construits manuellement à partir de ressources externes : le lexique des affects *LIDILEM* de (Augustyn *et al.*, 2006), une traduction du lexique *ANEW* de (Bradley & Lang, 1999) et un lexique d'émoticônes.

Le lexique *LIDILEM* modélise « le ressenti ou l'attitude du narrateur et/ou de ses personnages [qui] sont caractérisés lexicalement, comme relevant de la joie, de la tristesse, etc. » (Augustyn *et al.*, 2006). Ce lexique est décomposé en trois parties : les verbes, les adjectifs, les noms. Le lexique *ANEW* est constitué de 2476 mots évalué selon trois critères : plaisir, excitation et dominance. Pour chaque critère de toutes les entrées du lexique, les auteurs ont associé un score correspondant à la moyenne et à l'écart type. Le lexique des émoticônes a été construit manuellement à partir de la liste des émoticônes de Wikipedia [3]. De cette liste, nous n'avons gardé que 40 classes pour un total de 218 émoticônes. Nous avons ignoré les classes dont les émoticônes n'apparaissaient pas dans un corpus de plus de 7,5 millions de tweets français [4]. Cette approche compile 25 traits.

4.3 Construction semi-automatique de lexiques spécialisés à partir du corpus fourni

Dans la mesure où l'expression d'une opinion ou d'un sentiment dépend fortement de son contexte, nous nous sommes intéressés à l'acquisition d'éléments lexicaux caractéristiques de ces expressions au sein même du corpus fourni.

Cette extraction est réalisée en 3 étapes. Tout d'abord, nous entraînons un modèle de classification SVM sur le voisinage des mots du corpus afin de reconnaître les marqueurs d'opinion des autres mots. Dans le modèle appris, les exemples positifs sont les contextes entourant un marqueur connu, dans un lexique de marqueurs d'opinion peu dépendants du domaine (*horrible*, *catastrophique*, *joyeux*, etc.). Cette liste initiale contient environ 200 mots de ce type. Les traits de classification caractérisant un contexte appris sont :

— la forme lemmatisée et le rôle grammatical des mots entourant le candidat dans une fenêtre de 7 mots
— un marqueur indiquant si le contexte contient une négation (*ne, pas, jamais*, etc.)
— un marqueur indiquant si le contexte contient un déictique (*je, me, franchement*, etc.)

3. https://en.wikipedia.org/wiki/List_of_emoticons
4. Ce corpus représente une semaine de collecte à partir du 26 mars 2015

Le lexique extrait est ensuite filtré manuellement, afin de réduire le bruit dû aux erreurs de prétraitement ou aux ambiguïtés inhérentes à l'aspect subjectif de cette classification. Enfin nous recherchons les formes dérivées des mots acquis (flexions, variations grammaticales, ajout ou suppression d'affixes) pour les ajouter à notre ressource.

Nous avons réalisé cette extraction pour chaque catégorie sémantique, produisant ainsi 18 lexiques dont la taille varie fortement en fonction de la taille en nombre de tweets de la catégorie (de 4 à 86 mots). Cette approche compile 185 traits.

4.4 Des lexiques typés construits automatiquement à partir d'amorces et sur un corpus tiers

Approche Une de nos contributions réside dans la proposition d'une approche pour construire un lexique conséquent représentatif de classes sémantiques à reconnaître à partir d'amorces lexicales manuellement définies. Le principe de cette approche consiste dans un premier temps à définir des *graines*, instances des classes à reconnaître, puis dans un second temps à rechercher itérativement des variantes de celles-ci dans un corpus d'apprentissage et à les rajouter aux graines déjà récoltées avant de rechercher de nouvelles variantes de celles-ci. Le processus itératif peut se terminer après convergence (quand il n'y a plus de variantes à découvrir) ou au bout d'un nombre d'itérations prédéfinis (nous avons arbitrairement fixé cette limite à 20).

Pour définir nos graines nous sommes partis des classes spécifiques d'une classe telles que décrites dans le projet `ucomp` [5] et le document de présentation de DEFT'2015 [6]. Nous avons décliné les classes sémantiques avec une étiquette grammaticale (nom, adjectif ou participe passé, verbe infinitif, autres formes de verbes conjugués, adverbe). Le genre et le nombre des instances étaient masculin et singulier. Nous avons dérivé à la main les instances dans les catégories autres que nominale et autres formes de verbes conjuguées. Cette dernière forme a été automatiquement construite en sélectionnant dans les variantes récoltées les formes verbales qui n'étaient pas des autres étiquettes grammaticales retenues. La classe `MÉPRIS-NOM` était par exemple représentée par les graines *mépris, dédain, dégoût* et *haine*.

Pour rechercher les variantes, nous avons exploité la technique de construction de vecteurs de mots offerte par `word2vec` (Mikolov *et al.*, 2013). Cette technique présente l'avantage de rapprocher des formes avérées dans le corpus, variantes orthographiques, morphologiques et lexicales. Suivant cette approche, nous avons constitué différents lexiques à partir de différents jeux de classes : un jeu de classes décrivait la polarité, un autre les classes émotionnelles fines et un autre les classes émotionnelles fines (en fusionnant les classes antonymes). Le lexique classé en polarité a bénéficié d'un post-traitement qui consistait à changer la classe d'une occurrence en fonction de sa distribution sur les tweets annotés dans le corpus d'entraînement de DEFT. Pour un terme donné, si la différence entre son nombre d'occurrences dans un tweet positif et dans un tweet négatif était inférieure au nombre d'occurrences de neutre, le terme était classé neutre. Sinon il était classé en positif ou négatif selon le nombre d'occurrence dans la classe majoritaire.

Réalisation Pour construire ce lexique, nous avons exploité le corpus de tweets français utilisé pour filtrer le lexique d'émôticones décrit en section 4.2. Nous l'avons prétraité linguistiquement (uniformisation de la casse, tokenization et suppression des tokens non alphabétiques et d'un seul caractère) et nettoyé (retrait des tweets doublons).

Quel que soit le jeu de classes sémantiques initiales, nos graines sont au nombre de 273. La taille des lexiques construits diffèrent selon le jeu de classes initial : le lexique en polarité compte 2 650 termes, celui en classes émotionnelles fines 9 631 et celui en classes émotionnelles fines avec fusion des antonymes 4 804. Un oeil critique sur le contenu des classes obtenues nous conduit à relever quelques erreurs de classement, la présence de termes ambigüs et bien sûr un problème de complétude. Mais l'essentiel des regroupements reste cohérent. De part notre précédé de recherche de variantes, le filtrage grammatical joue un rôle primordial dans la qualité des lexiques extraits. Cette approche compile 128 traits.

4.5 Divers traits surfaciques et linguistiques

Nous avons défini des traits pour caractériser la forme des tweets. Parmi ces traits, nous comptons : le hashtag, le cashtag, la mention, l'url, le token entièrement (ou partiellement, ou débutant) en majuscules, le token constitué entièrement ou contenant au moins un symbole, le token constitué entièrement ou contenant au moins un chiffre, la répétition de caractères quelconques ou alphabétique ou numérique, de marques de ponctuation, le nombre de tokens. Nous comptons également un trait pour représenter chacun des lexiques fermés manuellement constitués suivants : déterminants et pronoms d'emphase, négation, comparaisons, pronoms selon leur personne. Cette approche compile 30 traits.

5. `http://www.ucomp.eu/`
6. `https://deft.limsi.fr/2015/descriptionTaches.fr.php?lang=fr`

5 Expériences et résultats

Après avoir brièvement présenté les données et le protocole expérimental, nous rapportons les scores de différentes approches sur le corpus d'entraînement et sur le corpus de test, à savoir l'approche de base, puis l'utilisation de lexiques exogènes et endogènes (`RUN 1`) à partir de cette approche de base, et enfin cette même approche mais en utilisant uniquement les lexiques endogènes (`RUN 2`). Nous discutons ensuite ces résultats.

5.1 Protocole expérimental et données

Par la suite, nous utilisons les mesures suivantes pour discuter de nos résultats. Un *vrai positif* est un test jugé correctement positif, un *faux positif* est un test incorrectement jugé positif, un *faux négatif* est un test incorrectement jugé négatif. La *précision* d'un système correspond au nombre de tests positifs corrects sur le nombre de tests estimés positifs (somme des corrects et des incorrects). Le *rappel* est le nombre de tests positifs corrects sur la nombre total de tests positifs réels (somme des vrais positifs et des faux négatifs). La *F-mesure* est une moyenne harmonique de la précision et du rappel. La *micro-précision* est le nombre de tests positifs corrects toute classe confondue (somme des vrais positifs quelle que soit la classe) sur le nombre de tests estimés positifs toute classe confondue. La *macro-précision* est la moyenne des précisions obtenues sur chaque classe. Le *macro-rappel* est la moyenne des rappels obtenus pour chaque classe.

En pratique, les mesures de précision utilisées dans DEFT sont légèrement différentes puisqu'elles comptabilisent un faux négatif à chaque classe dès qu'une instance n'a pas été classée dans une des classes disponibles. Cette situation est rencontrée à chaque fois qu'un système a assigné une instance à la classe `INFORMATION`. Dans la mesure classique, seuls les faux positifs de cette classe sont incrémentés. Pour les organisateurs, cette approche vise à "fortement pénaliser les systèmes qui ne répondent pas (ou qui fournissent des classes non prévues) alors que c'était une information disponible". De cette manière, un système qui attribue une classe prévue pour un tweet mais se trompe aura un meilleur score qu'un système qui ne choisit pas ou qui "invente" une classe.

Les scores sur le corpus d'entraînement ont été obtenus par validation croisée sur 10 partitions ; les moyennes des scores sont obtenues par 10 systèmes entraînés sur 9 partitions et testés sur la dixième. Le contenu des partitions a initialement été tiré au hasard. Le corpus d'entraînement compte 6 672 instances dont 3 531 classées `INFORMATION` et 3 142 `NON-INFORMATION`. D'après ce corpus, un système qui classerait toutes les instances dans cette classe majoritaire aurait, pour la tâche 2.2, 0.5292 de micro-précision et 0.0294 de macro-précision. Le corpus de test compte lui 3 379 instances pour les tâches 1 et 2.1 avec 1 861 `INFORMATION` et 1 518 `NON-INFORMATION`. Le corpus de test pour la tâche 2.2 compte 1 361 instances `NON-INFORMATION`. Pour la tâche 1, nous n'avons pas considéré la classe mixte. Pour la tâche 2.2, nous avons considéré la classe `INFORMATION` comme étant une des classes fines à reconnaître en plus des 18 autres définies.

Classe	**Approche de base**			**RUN 1**			**RUN 2**		
	Mic-P	**Mac-P**	**Mac-R**	**Mic-P**	**Mac-P**	**Mac-R**	**Mic-P**	**Mac-P**	**Mac-R**
Polarité (t1)	0.5969	0.6647	0.5224	**0.714**	**0.7582**	**0.619**	0.6736	0.6783	0.6021
Méta-classe (t2.1)	0.5634	0.4876	0.3930	**0.71**	**0.702**	0.438	0.6649	0.6231	**0.4865**
Classe fine (t2.2)	0.2792	0.0217	0.0224	**0.681**	**0.518**	0.241	0.6349	0.4709	**0.2604**

Tableau 1 – Performance de l'approche de base, `RUN 1` et `RUN 2` sur le corpus d'*entraînement* avec Micro-Précision, Macro-Précision et Macro-Rappel.

Classe	**Général**					**Approche de base**		**RUN 1**		**RUN 2**	
	Moy	Méd	E-T	Min	Max	**Mic-P**	**Mac-P**	**Mic-P**	**Mac-P**	**Mic-P**	**Mac-P**
Polarité (t1)	0.581	0.693	0.238	0.04	0.735	0.5969	0.6647	0.6087	0.6552	**0.6232**	**0.6769**
Méta-classe (t2.1)	0.408	0.514	0.217	0.029	0.612	0.5634	0.4876	0.5711	0.5081	**0.5750**	**0.5143**
Classe fine (t2.2)	0.179	0.199	0.152	0	0.346	0.2792	0.0217	**0.3343**	**0.0281**	0.3159	0.0273

Tableau 2 – Performance de l'approche de base, `RUN 1` et `RUN 2` sur le corpus de *test* avec Moyenne, Médiane, Ecart-type, Min, Max, Micro-Précision et Macro-Précision.

5.2 Discussion des résultats

Le tableau 1 rapporte les résultats globaux obtenus sur le corpus d'entraînement et le tableau 2 sur le corpus de test ; avec dans les deux cas les mesures d'évaluation modifiées dans le cadre de DEFT. Dans le tableau 1, les résultats de l'approche `RUN 1` ont été calculés sur un échantillonnage en 10 parties et ceux de l'approche `RUN 2` sur un échantillonnage en 2 parties. Pour la suite de l'analyse, nous supposons que cela n'affecte pas les résultats.

Les tableaux 3, 4 et 5 présentent le détail des scores obtenus par l'approche de base sur les différentes classes avec les mesures d'évaluation originales sur le corpus d'entraînement. Les tableaux 6, 7 et 8 présentent le détail des scores obtenus par l'approche `RUN 1` sur les différentes classes avec les mesures d'évaluation originales sur le corpus d'entraînement. Les tableaux 9, 10 et 11 présentent le détail des scores obtenus par l'approche `RUN 2` sur les différentes classes avec les mesures d'évaluation originales sur le corpus d'entraînement.

Les résultats présentés dans le tableau 1 indiquent que les deux approches proposées sont plus efficaces que l'approche de base sur le corpus d'entraînement. L'approche `RUN 1` obtient une meilleure micro et macro précisions que l'approche `RUN 2`, mais cette dernière obtient un meilleur rappel.

Sur le corpus de test, les deux approches proposées sont également meilleures que l'approche de base (tableau 2). Cependant, l'approche `RUN 1` n'est meilleure que sur la tâche 2.2. L'approche `RUN 2` obtient les meilleures performances sur les tâches 2.1 et 1.

Après analyse des résultats, nous pensons que l'approche `RUN 2` est meilleure sur les tâches 2.1 et 1 car, même si l'approche `RUN 1` obtient de meilleurs résultats sur les classes les plus représentées. Cette différence tend à s'estomper quand les classes sont généralisées. Nous remarquons que les scores de rappel de l'approche `RUN 2` sont globalement meilleurs sur la classe générique `ÉMOTION`. Or cette classe générique regroupe 12 classes ce qui représente un fort déséquilibre par rapport aux classes génériques `OPINION` et `SENTIMENT` qui ne regroupent que 6 classes à elles deux. Les scores de rappel important de l'approche `RUN 2` sur les classes de l'`ÉMOTION` se traduisent par une précision accrue lorsque les classes sont regroupées au sein des classes génériques.

À la lumière de ces résultats, nous constatons que les lexiques endogènes (approche `RUN 2`) permettent de faire remonter des tweets de la classe générique `ÉMOTION` au détriment de la précision sur l'ensemble des classes. À la différence de l'approche `RUN 2` dont la combinaison de lexiques endogènes et exogènes permet au contraire de désambiguïser un grand nombre de classe au détriment des moins représentées.

6 Conclusion et perspectives

Dans ce défi en fouille de texte, il s'agit de retrouver les classes sémantiques de tweets, à différents niveaux de granularité. Nous choisissons de réaliser une classification des tweets au niveau le plus fin à l'aide d'une représentation en bigrammes de mots et de plusieurs lexiques affectifs, puis d'inférer les classes sémantiques plus générales selon cette première classification. Les résultats de cette inférence sont globalement meilleurs que ceux de la classification fine, ce qui peut s'expliquer par la faible ambiguïté entre les classes sémantiques générales comparativement aux plus fines. Toutefois, les erreurs de classification au niveau le plus fin ont nécessairement des répercussions sur les classes générales. En effet, quelques classes fines sémantiquement proches (les classes `DÉSACCORD` et `DÉPLAISIR` par exemple) mais ne partagent pas la même classe d'information (respectivement, `OPINION` ou `ÉMOTION`). En ce qui concerne la classification fine des tweets, nous constatons que l'approche de base utilisant une représentation en sac de bigrammes de mots fournit des résultats satisfaisants sur les classes les plus présentes, mais ne parvient pas à désambiguïser avec justesse les classes moins fréquentes. Les lexiques affectifs que nous avons utilisés permettent d'affiner dans une certaine mesure cette classification. Nous observons cependant deux tendances selon le type de lexique employé. Dans le cas des lexiques affectifs construits indépendamment du corpus fourni, la plupart des classes bénéficient d'un gain modéré tandis qu'en utilisant les lexiques issus du corpus de ce défi, quelques classes sémantiques sont particulièrement bien identifiées, au détriment des autres.

Ce travail a été l'occasion de nous intéresser à la catégorisation d'émotions pour la fouille d'opinion, et nous envisageons de poursuivre certaines pistes dans ce domaine. Parmi celles-ci, nous prévoyons de comparer l'intérêt du point de vue de la désambiguïsation de l'opinion des expressions ou multi-mots aux unigrammes de mots et de mesurer l'apport de l'analyse syntaxique pour une telle tâche.

Classe	ref.	hyp.	correct	Précision	Rappel	F-Mesure
+	1772	1262	846	0.6703	0.4774	0.5576
=	3531	4727	3224	0.6820	0.9130	0.7808
-	1370	684	561	0.8201	0.4094	0.5462
Total	6673		4631	0.6939 (Micro-P)	0.5999 (Macro-R)	
				0.7241 (Macro-P)		

Tableau 3 – Détail de la performance de l'*approche de base* pour la *tâche 1* sur le corpus d'*entraînement* avec nombre absolu en référence, dans l'hypothèse et dans l'ensemble correct ainsi que les scores de Précision, Rappel F-mesure, Micro-Précision et Macro-Précision originale.

Classe	ref.	hyp.	correct	Précision	Rappel	F-Mesure
EMOTION	776	354	264	0.7457	0.3402	0.4672
INFORMATION	3531	4727	3224	0.6820	0.9130	0.7808
OPINION	2250	1551	1107	0.7137	0.492	0.5824
SENTIMENT	82	40	26	0.65	0.3170	0.4262
Total	6673		4621	0.6924 (Micro-P)	0.4124 (Macro-R)	
				0.5583 (Macro-P)		

Tableau 4 – Détail de la performance de l'*approche de base* pour la *tâche 2.1* sur le corpus d'*entraînement* avec nombre absolu en référence, dans l'hypothèse et dans l'ensemble correct ainsi que les scores de Précision, Rappel F-mesure, Micro-Précision et Macro-Précision originale.

Classe	ref.	hyp.	correct	Précision	Rappel	F-Mesure
ACCORD	153	47	31	0,6596	0,2026	0,3100
AMOUR	8	0	0	0,0000	0,0000	0,0000
APAISEMENT	9	2	0	0,0000	0,0000	0,0000
COLERE	206	106	67	0,6321	0,3252	0,4295
DEPLAISIR	47	1	0	0,0000	0,0000	0,0000
DERANGEMENT	12	5	5	1,0000	0,4167	0,5882
DESACCORD	212	132	97	0,7348	0,4575	0,5640
DEVALORISATION	394	202	106	0,5248	0,2690	0,3557
ENNUI	4	0	0	0,0000	0,0000	0,0000
INFORMATION	3531	4727	3224	0,6820	0,9131	0,7808
INSATISFACTION	9	3	2	0,6667	0,2222	0,3333
MEPRIS	173	42	8	0,1905	0,0462	0,0744
PEUR	269	191	155	0,8115	0,5762	0,6739
PLAISIR	34	6	3	0,5000	0,0882	0,1500
SATISFACTION	73	37	24	0,6486	0,3288	0,4364
SURPRISE_NEGATIVE	10	1	1	1,0000	0,1000	0,1818
SURPRISE_POSITIVE	4	0	0	0,0000	0,0000	0,0000
TRISTESSE	34	1	0	0,0000	0,0000	0,0000
VALORISATION	1491	1170	708	0,6051	0,4748	0,5321
Total	6673		4431	0,6640 (Micro-P)	0,2327 (Macro-R)	
				0,4556 (Macro-P)		

Tableau 5 – Détail de la performance de l'*approche de base* pour la *tâche 2.2* sur le corpus d'*entraînement* avec nombre absolu en référence, dans l'hypothèse et dans l'ensemble correct ainsi que les scores de Précision, Rappel F-mesure, Micro-Précision et Macro-Précision originale.

Classe	ref.	hyp.	correct	Précision	Rappel	F-Mesure
+	1772	1156	836	0,7232	0,4718	0,5710
=	3531	4808	3319	0,6903	0,9400	0,7960
-	1370	709	611	0,8618	0,4460	0,5878
Total	6673		4766	0,7142 (Micro-P) 0,75841 (Macro-P)	0,6192 (Macro-R)	

Tableau 6 – Détail de la performance du `RUN 1` pour la *tâche 1* sur le corpus d'*entraînement* avec nombre absolu en référence, dans l'hypothèse et dans l'ensemble correct ainsi que les scores de Précision, Rappel F-mesure, Micro-Précision et Macro-Précision originale.

Classe	ref.	hyp.	correct	Précision	Rappel	F-Mesure
EMOTION	776	395	318	0,8051	0,4098	0,5431
INFORMATION	3531	4808	3319	0,6903	0,9400	0,7960
OPINION	2250	1434	1083	0,7552	0,4813	0,5879
SENTIMENT	82	29	20	0,6897	0,2439	0,3604
Total	6673		4744	0,7109 (Micro-P) 0,7023 (Macro-P)	0,4385 (Macro-R)	

Tableau 7 – Détail de la performance du `RUN 1` pour la *tâche 2.1* sur le corpus d'*entraînement* avec nombre absolu en référence, dans l'hypothèse et dans l'ensemble correct ainsi que les scores de Précision, Rappel F-mesure, Micro-Précision et Macro-Précision originale.

Classe	ref.	hyp.	correct	Précision	Rappel	F-Mesure
ACCORD	153	65	40	0,6154	0,2614	0,3670
AMOUR	8	0	0	0,0000	0,0000	0,0000
APAISEMENT	9	1	0	0,0000	0,0000	0,0000
COLERE	206	107	72	0,6729	0,3495	0,4601
DEPLAISIR	47	1	0	0,0000	0,0000	0,0000
DERANGEMENT	12	4	4	1,0000	0,3333	0,5000
DESACCORD	212	134	96	0,7164	0,4528	0,5549
DEVALORISATION	394	176	103	0,5852	0,2614	0,3614
ENNUI	4	0	0	0,0000	0,0000	0,0000
INFORMATION	3531	4808	3319	0,6903	0,9400	0,7960
INSATISFACTION	9	2	2	1,0000	0,2222	0,3636
MEPRIS	173	47	18	0,3830	0,1040	0,1636
PEUR	269	230	182	0,7913	0,6766	0,7295
PLAISIR	34	4	2	0,5000	0,0588	0,1053
SATISFACTION	73	27	18	0,6667	0,2466	0,3600
SURPRISE_NEGATIVE	10	1	1	1,0000	0,1000	0,1818
SURPRISE_POSITIVE	4	0	0	0,0000	0,0000	0,0000
TRISTESSE	34	7	4	0,5714	0,1176	0,1951
VALORISATION	1491	1059	688	0,6497	0,4614	0,5396
Total	6673		4549	0,6817 (Micro-P) 0,5180 (Macro-P)	0,2414 (Macro-R)	

Tableau 8 – Détail de la performance du `RUN 1` pour la *tâche 2.2* sur le corpus d'*entraînement* avec nombre absolu en référence, dans l'hypothèse et dans l'ensemble correct ainsi que les scores de Précision, Rappel F-mesure, Micro-Précision et Macro-Précision originale.

Classe	ref.	hyp.	correct	Précision	Rappel	F-Mesure
+	1771	1261	765	0.6066	0.4319	0.5046
=	3530	4463	3016	0.6757	0.8543	0.7546
-	1369	946	712	0.7526	0.5200	0.6151
Total	6670		4493	0.6736 (Micro-P) 0.6783 (Macro-P)	0.6021 (Macro-R)	

Tableau 9 – Détail de la performance du RUN 2 pour la *tâche 1* sur le corpus d'*entraînement* avec nombre absolu en référence, dans l'hypothèse et dans l'ensemble correct ainsi que les scores de Précision, Rappel F-mesure, Micro-Précision et Macro-Précision originale.

Classe	ref.	hyp.	correct	Précision	Rappel	F-Mesure
EMOTION	809	463	342	0.7386	0.4227	0.5377
INFORMATION	3530	4463	3016	0.6757	0.8543	0.7546
OPINION	2250	1709	1061	0.6208	0.4715	0.5359
SENTIMENT	81	35	16	0.4571	0.1975	0.2758
Total	6670		4435	0.6649 (Micro-P) 0.6231 (Macro-P)	0.4865 (Macro-R)	

Tableau 10 – Détail de la performance du RUN 2 pour la *tâche 2.1* sur le corpus d'*entraînement* avec nombre absolu en référence, dans l'hypothèse et dans l'ensemble correct ainsi que les scores de Précision, Rappel F-mesure, Micro-Précision et Macro-Précision originale.

Classe	ref.	hyp.	correct	Précision	Rappel	F-Mesure
ACCORD	153	70	38	0.5428	0.2483	0.3408
AMOUR	8	1	0	0.0	0.0	0
APAISEMENT	9	3	2	0.6666	0.2222	0.3333
COLERE	206	72	52	0.7222	0.2524	0.3741
DEPLAISIR	47	6	0	0.0	0.0	0
DERANGEMENT	12	3	3	1.0	0.25	0.4
DESACCORD	212	163	105	0.6441	0.4952	0.56
DEVALORISATION	394	331	156	0.4712	0.3959	0.4303
ENNUI	4	0	0	0	0.0	0
INFORMATION	3530	4463	3016	0.6757	0.8543	0.7546
INSATISFACTION	9	9	0	0.0	0.0	0
MEPRIS	172	89	36	0.4044	0.2093	0.2758
PEUR	269	258	191	0.7403	0.7100	0.7248
PLAISIR	34	16	7	0.4375	0.2058	0.28
SATISFACTION	72	26	16	0.6153	0.2222	0.3265
SURPRISE_NEGATIVE	10	3	3	1.0	0.3	0.4615
SURPRISE_POSITIVE	4	0	0	0	0.0	0
TRISTESSE	34	12	6	0.5	0.1764	0.2608
VALORISATION	1491	1146	605	0.5279	0.4057	0.4588
Total	6670		4236	0.6349 (Micro-P) 0.4709 (Macro-P)	0.2604 (Macro-R)	

Tableau 11 – Détail de la performance du RUN 2 pour la *tâche 2.2* sur le corpus d'*entraînement* avec nombre absolu en référence, dans l'hypothèse et dans l'ensemble correct ainsi que les scores de Précision, Rappel F-mesure, Micro-Précision et Macro-Précision originale.

Remerciement

Ce travail a bénéficié du soutien du fond unique interministériel (FUI) 17 au travers du projet ODISAE [7] ainsi que d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01 (project LIMAH [8]).

Références

AUGUSTYN M., BEN HAMOU S., BLOQUET G., GOOSSENS V., LOISEAU M. & RINCK F. (2006). Lexique des affects : constitution de ressources pédagogiques numériques. In *Colloque International des étudiants-chercheurs en didactique des langues et linguistique.*, Grenoble, France.

BRADLEY M. M. & LANG P. J. (1999). Affective norms for english words (ANEW) : Instruction manual and affective ratings.

CRAMMER K. & SINGER Y. (2000). On the learnability and design of output codes for multiclass problems. In *Proceedings of COLT '00*, p. 35–46, Palo Alto, CA, USA.

FAN R.-E., CHANG K.-W., HSIEH C.-J., WANG X.-R. & LIN C.-J. (2008). LIBLINEAR : a library for large linear classification. *Journal of Machine Learning Research*, **9**, 1871–1874.

HSU C.-W., CHANG C.-C. & LIN C.-J. (2003). *A practical guide to support vector classification*. Rapport interne, Department of Computer Science, National Taiwan University.

LIU B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, **5**(1).

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, **abs/1310.4546**.

MURRAY G. & CARENINI G. (2011). Subjectivity detection in spoken and written conversations. *Natural Language Engineering*, **1**(1).

PARROTT W. (2001). *Emotions in Social Psychology*. Philadelphia, PA, USA : Psychology Press.

PUSTEJOVSKY J. & STUBBS A. (2012). *Natural Language Annotation for Machine Learning*. O'Reilly Publishers.

STAIANO J. & GUERINI M. (2014). DepecheMood : a Lexicon for Emotion Analysis from Crowd-Annotated News. *CoRR*, p. 427–433.

VERNIER M., MONCEAUX L., DAILLE B. & DUBREIL E. (2009). Catégorisation des évaluations dans un corpus de blogs multi-domaine. *Revue des Nouvelles Technologies de l'Information (RNTI)*.

WIEBE J. & MIHALCEA R. (2006). Word sense and subjectivity. In *Proceedings of ACL'06*, Sydney, Australia.

YANG M., PENG B., CHEN Z., ZHU D. & CHOW K. (2014). A Topic Model for Building Fine-grained Domain-specific Emotion Lexicon. *anthology.aclweb.org*, p. 421–426.

YU H.-F., HO C.-H., JUAN Y.-C. & LIN C.-J. (2013). *LibShortText : a library for short-text classification and analysis*. Rapport interne, Department of Computer Science, National Taiwan University. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libshorttext`.

7. www.odisae.com
8. limah.irisa.fr