Construction et exploitation d'un corpus français pour l'analyse de sentiment

Marc Vincent[1] Grégoire Winterstein[2]
(1) UMR S-775, Université Paris Descartes
(2) LLF, UMR 7110, Université Sorbonne Nouvelle
marc.r.vincent@gmail.com, gregoire.winterstein@linguist.jussieu.fr

RÉSUMÉ

Ce travail présente un corpus en français dédié à l'analyse de sentiment. Nous y décrivons la construction et l'organisation du corpus. Nous présentons ensuite les résultats de l'application de techniques d'apprentissage automatique pour la tâche de classification d'opinion (positive ou négative) véhiculée par un texte. Deux techniques sont utilisées : la régression logistique et la classification basée sur des *Support Vector Machines* (SVM). Nous mentionnons également l'intérêt d'appliquer une sélection de variables avant la classification (par régularisation par *elastic net*).

ABSTRACT

Building and exploiting a French corpus for sentiment analysis

This work introduces a French corpus for sentiment analysis. We describe the construction and organization of the corpus. We then apply machine learning techniques to automatically predict whether a text is positive or negative (the opinion classification task). Two techniques are used : logistic regression and classification based on *Support Vector Machines* (SVM). Finally, we briefly evaluate the merits of applying feature selection algorithms to our models (via elastic net regularization).

MOTS-CLÉS : Analyse de sentiments, Corpus, Classification, Apprentissage automatique, Sélection de variable.

KEYWORDS: Sentiment Analysis, Corpus, Opinion Mining, Classification, Machine Learning, Variable Selection.

1 Introduction

Ce travail présente la construction et l'exploitation d'un corpus français destiné à l'analyse de sentiment (*sentiment analysis* ou *opinion mining*). L'analyse de sentiment recouvre l'ensemble des tâches dédiées à la reconnaissance des opinions exprimées au sein d'un texte et connaît de nombreuses applications (pour un panorama voir Pang et Lee (2008)).

Les recherches sur l'analyse de sentiments (ou de subjectivité) sont en majorité centrées sur l'anglais, bien que le sujet ait déjà fait l'objet de plusieurs recherches ayant abouti entre autres à l'établissement de corpus (cf. notamment Grouin et al. (2007); Vernier (2011)). Nous avons cependant jugé utile de construire un nouveau corpus constitué de critiques issues du web. La motivation, la construction et la structure du corpus sont décrites en section 2.

Parmi les tâches relevant de l'analyse de sentiment nous nous sommes focalisés sur la classification d'opinion, c'est-à-dire sur la tâche qui consiste à classer un texte dans une catégorie d'opinion (typiquement *positif* ou *négatif*). Nous rapportons les résultats obtenus pour cette tâche en ayant eu recours aux *Support Vector Machines* (SVM). Nous rapportons également les résultats obtenus en opérant au préalable une sélection des variables par régularisation par *elastic net*. Notre méthodologie est décrite des sections 3.1 à 3.4 et nos résultats en section 3.5.

2 Construction et constitution du corpus

Les ressources nécessaires à l'analyse de sentiment doivent fournir en parallèle d'un contenu textuel une forme d'évaluation du sentiment associé au texte. Avec le développement des contenus générés par les utilisateurs sur le web, ce type de ressource peut aujourd'hui facilement s'obtenir sur des sites web permettant aux internautes de partager leur opinion sur divers sujets. Du point de vue qualitatif et méthodologique, un corpus regroupant ce genre de textes se doit d'être le plus général possible afin que les modèles issus de techniques d'apprentissage soient aussi généraux que possible. Cela signifie notamment que chacune des critiques récupérées doit traiter d'un produit différent. Les descriptions des corpus existants (p.ex. celui utilisé dans la campagne DEFT'07 ou celui utilisé par Ghorbel et Jacot (2011)) ne font pas état de la variété d'éléments évalués dans le corpus, et il nous a donc paru pertinent de construire un nouveau corpus en tenant compte de cette dimension.

2.1 Construction du corpus

La construction de notre corpus s'appuie sur la collecte de commentaires d'internautes recueillis sur différentes plate-formes web en français et permettant aux utilisateurs d'exprimer leur opinion par le biais d'une note chiffrée. La totalité des informations a été obtenue de manière automatique et non-supervisée en créant des parseurs adaptés aux sites concernés. Le corpus obtenu est disponible sur demande auprès des auteurs.

Afin de varier les thèmes abordés dans les critiques constituant le corpus, nous avons considéré trois domaines différents : des critiques de films tirées du site `allocine.fr`, des avis sur des romans de poche extraits du site `amazon.fr` et des commentaires relatifs à des établissements hôteliers tirés du site `tripadvisor.fr`. Sur chacun de ces sites, les utilisateurs sont invités à rédiger une opinion et à exprimer leur avis par une note située entre 1 et 5 (typiquement représentée à l'écran par un nombre d'étoiles). Le nombre de commentaires par type de produit est résumé dans le tableau ci-dessous :

Type de produit	Provenance	N. critiques / note
Hôtels	`tripadvisor.fr`	1000
Films	`allocine.fr`	1000
Romans	`amazon.fr`	800

TABLE 1 – Constitution du corpus (nombre total de textes : 14000)

La diversité de provenance des critiques est une première étape pour s'assurer que les modèles

produits par les algorithmes d'apprentissage ne se montreront pas trop liés au idiosyncrasies du corpus. Outre la diversité de provenance, nous nous sommes également assurés que :

– Le nombre de produits distincts représentés dans chaque plage de notation soit maximal. Dans le cas des romans et des films, ce nombre est égal au nombre de critiques, c'est-à-dire qu'un produit donné fait l'objet d'au plus une critique pour chacune des 5 notes. Dans le cas des établissements hôteliers cette ventilation ne s'est pas montrée possible, nous avons donc cherché à limiter le nombre de répétitions. Au final, aucun produit ne se trouve mentionné plus de 5 fois par plage de note dans le corpus.
– Les auteurs de critiques soient aussi différents que possible au sein des critiques d'une même plage de notation. S'il n'a pas été possible de s'assurer que chacune des critiques ait un auteur distinct des autres, le nombre de critiques signées d'un même auteur au sein d'une même plage de notation est de 12. [1]

Ces deux précautions permettent d'éviter que des modèles produits par des algorithmes d'apprentissage perdent en généralité en étant trop dépendant des spécificités propres à certains items ou auteurs fréquemment répétés.

Les possibilités de notation offertes par les plates-formes marchandes vont aujourd'hui au-delà de l'appariement d'une note à un texte. C'est pourquoi, outre la note attribuée et le contenu du commentaire utilisateur, nous avons cherché à conserver la totalité des informations disponibles et pertinentes pour différentes tâches d'analyse de sentiment. Chacune des critiques utilisateurs est alors accompagnée des informations suivantes (le corpus est organisé dans un format `XML`) : [2]

– Identifiant de la critique.
– Identifiant du produit (sous forme d'entier).
– Descriptif du produit (type de produit et titre de film, nom de l'hôtel ou titre de roman).
– Note associée à la critique (fournie par l'auteur de la critique).
– Identifiant (anonymisé) de l'auteur de la critique
– Contenu de la critique

Dans les parties relatives à `tripadvisor` et `amazon` on trouve de plus pour chaque critique individuelle :

– Le résumé de la critique (en une phrase) fourni par l'utilisateur.
– Une mesure "d'utilité" de la critique, indiquée par le nombre d'utilisateurs ayant jugé la critique utile.

Enfin, dans la partie des critiques issues de `tripadvisor`, on inclut également des informations de notation sur des critères spécifiques. Par exemple certains utilisateurs notent la propreté des chambres ou le rapport qualité/prix de l'hôtel (toujours sur une note de 1 à 5).

Au final, la longueur totale du corpus, en nombre de termes différents reconnus par segmentation automatique (en utilisant le tokenizer de `MElt`, cf. infra) est de 1 402 867 tokens. La longueur moyenne d'une critique est de 100 tokens, les critiques d'établissements hôteliers se montrant globablement plus longues (123 tokens en moyenne) que celles de films (90 tokens) ou de romans (83 tokens).

1. La plage de notation en question est celle correspondant à une note de 2 sur 5 pour les critiques de films. Cette plage de notation s'avère sévèrement sous-représentée de manière générale sur le site `allocine.fr`. En excluant cette plage spécifique, le nombre maximal de répétitions par auteur dans le corpus est de 5.

2. Les informations récupérées n'ont pas fait l'objet de validation subséquente, notamment sur l'adéquation des notes indiquées par les utilisateurs avec le contenu de leur critique. Cependant, nous avons effectué une extraction aléatoire de 150 critiques notées 1 ou 5 que nous avons manuellement annotées en "positif" et "négatif". Sur les 150, une seule erreur a été relevée, montrant que les données utilisées ultérieurement dans la tâche de classification sont fiables.

3 Classification d'opinion par apprentissage automatique

En guise d'illustration de l'emploi du corpus construit, nous nous focalisons sur la tâche de classification automatique d'opinion. Cette tâche est une des premières qui ait été abordée dans le domaine de l'*opinion mining* (Pang *et al.*, 2002) et il nous est apparu pertinent de fournir un étalon relatif au corpus que nous utilisons. Il est important de noter que cette tâche ne fait pas appel à la totalité des informations offertes par le corpus : d'autres tâches potentiellement plus complexes peuvent par exemple faire usage des indicateurs d'utilité associés aux critiques. Un de nos buts est avant tout de fournir des mesures de bases associées au corpus. Comme mentionné précédemment, des tâches similaires ont déjà été entreprises, mais sur des corpus dont le caractère général n'est pas assuré. Outre de fournir ces mesures, nous voulons également mesurer l'intérêt d'appliquer des techniques de réduction de variable avant les phases d'apprentissage automatique.

Pour aborder la tâche de classification automatique nous avons utilisé deux types d'approches : une classification basée sur des SVM, et une autre basée sur la régression logistique à l'issue de la sélection de variable. Pour chacune de ces tâches, les critiques ont été au préalable segmentée, étiquetée et lemmatisée. Du fait du marquage morphologique relativement riche du français cette étape est apparue nécessaire pour optimiser les performances des modèles produits. Pour cette étape nous nous sommes basés sur l'étiqueteur et lemmatiseur `MElt` (Denis et Sagot, 2012) dont les performances pour le français sont l'état de l'art et qui emploie un module de gestion de "crappy text" qui permet de corriger un certain nombre des irrégularités typiquement trouvées dans les contenus récupérés sur le web.

3.1 Traits

Pang *et al.* (2002) ont observé que pour la tâche de classification d'opinion, la façon la plus efficace de décrire un texte était sous forme de "sac de mots", c'est-à-dire d'un vecteur à valeurs booléennes, indiquant la présence et l'absence d'un élément lexical. Cette méthode s'avère plus efficace que d'encoder le nombre d'occurrences de chaque unigramme et meilleure que d'utiliser une description en termes de combinaisons d'unigrammes et de bigrammes.

Nous avons suivi ici leurs recommandations pour encoder nos données. Suite au processus de lemmatisation précédemment mentionné, nous avons retenu uniquement les lemmes qui avaient été reconnus par `MElt` et nous avons encodé chacune des critiques sous forme d'un vecteur encodant l'absence ou la présence d'un lemme (repéré par une combinaison de forme et de partie du discours, p.ex. `andalou/ADJ`). Nous avons sciemment omis de considérer les éléments non reconnus et ceux catégorisés comme noms propres : la prise en compte de ces éléments aurait fait baisser la généralité des modèles produits et nettement augmenté la taille de l'espace des traits (sur le corpus entier on dénombre 12 300 traits ainsi retenus contre 26 765 si tous les lemmes avaient été pris en compte).

La prise en compte de la présence d'une négation est traditionnellement considérée comme un indicateur pertinent pour la classification d'opinion. Usuellement, cette prise en compte se fait sous la forme d'un dédoublement des lemmes pris en compte : si un lemme se trouve sous la portée d'une négation dans la phrase, il sera encodé sous la forme d'un trait `NEG-LEMME`. Afin de mesurer l'impact de cette prise en compte, nous n'avons pas inclus la négation dans nos

expériences de base, et en évaluons séparement l'intérêt. L'algorithme de détection de la présence de négation est basique et en partie inspiré de celui de Das et Chen (2001). Afin d'étiqueter un lemme négativement nous avons :

– utilisé un chunk parser minimal (implémenté en `nltk`) pour identifier des structures négatives (p.ex. des verbes accompagnés d'un marqueur de négation),
– étiqueté tout élément situé à droite de la frontière gauche de ces constituants comme négatif.

On pourrait raffiner cette approche en utilisant un analyseur syntaxique en dépendances ou bien en utilisant un lexique de polarité (sur le modèle de Wilson *et al.* (2005)), mais nous réservons ces manipulations à une recherche future.

3.2 La tâche de classification d'opinion

Pour la première exploitation du corpus, nous avons choisi une tâche simple de classification d'opinion dans la lignée de celle entreprise par Pang *et al.* (2002). Le but de la tâche est donc de classer des documents (les critiques de produits) selon la polarité de l'opinion qui y est exprimée : positive ou négative. Pour les besoins de cette tâche, nous n'avons considéré que deux types de critiques : celles ayant reçu une note de 1 que nous avons considérées comme négatives, et celles ayant reçu une note de 5 que nous avons considérées comme positives. Nous n'avons pas chercher à classer individuellement les phrases qui composent chacun des documents selon leur polarité, bien que ce type de traitement permette généralement d'obtenir de meilleurs résultats (Pang et Lee, 2008).

Comme déjà mentionné nous avons utilisé deux techniques d'apprentissage automatique : une classification basée sur la régression logistique (en utilisant le package *R glmnet*) et une classification basée sur les *Support Vector Machines* (*SVM*). Pour cette dernière approche nous avons utilisé le programme SVM^{light} (Joachims, 1999). Le choix de ces méthodes est motivé par l'efficacité reconnue des méthodes SVM d'une part, et la simplicité de la régression logistique d'autre part.

3.3 Sélection de modèles et évaluation

Afin de reporter des estimateurs de performances non biaisés et réalistes des classifieurs testés, nous avons eu recours à une procédure de validation croisée imbriquée. Suivant le principe de la validation croisée, l'ensemble des N exemples est divisé aléatoirement en k partitions de test de même taille (N/k) et de même stratification (comportant le même nombre d'exemples de chaque classe et de chaque source). Chaque partition de test sert à l'évaluation d'un classifieur construit à partir du reste des exemples du corpus (de taille $N - N/k$) qui forme la partition d'apprentissage.

Comme nous utilisons des algorithmes d'apprentissage paramétrés (SVM et elastic net) nous devons au préalable avoir déterminé les valeurs de ces paramètres par une validation croisée interne qui, à partir de chacune de k partitions d'apprentissage, crée m partitions de test et d'apprentissage. Les paramètres sélectionnés parmi ceux testés sont ceux qui auront permis d'obtenir la meilleure moyenne d'une mesure de performance (déviance pour l'elastic net, F-score pour les SVM) mesurée pour chaque partition de test de la validation croisée interne. Pour nos expériences nous avons choisi $k = 10$ et $m = 5$. En préalable à ces expériences nous avons filtré les variables présentes dans moins de 10 exemples dans le corpus afin de faciliter l'apprentissage

et l'interprétation des modèles produits (aboutissant à 2829 variables sans `NEG-LEMME`, 3257 avec).

3.4 Sélection de variable

Les techniques de sélection de variables visent à réduire le nombre de traits utilisés dans les modèles produits par les techniques d'apprentissage automatique. La sélection de variable poursuit trois objectifs reliés entre eux (Guyon et Elisseeff, 2003) :
– L'amélioration de la performance des modèles prédicteurs.
– La mise au point de prédicteurs plus rapides et consommant moins de ressources.
– Une meilleure compréhension des processus à l'œuvre dans la génération des données.
Il existe plusieurs méthodes de sélection de variable. Pour nos expériences nous avons utilisé la régularisation par *elastic net* sur un modèle de régression logistique (Zou et Hastie, 2005) (à l'aide de la bibiliothèque `glmnet`) pour deux raisons. D'une part, à l'instar d'autres techniques lomme le LASSO, l'*elastic net* produit des modèles creux en éliminant les variables non essentielles à la prédiction, cependant l'*elastic net* inclut dans le modèle l'ensemble des variables prédictives même lorsque celle-ci sont corrélées entre elles (alors que le LASSO tend à n'en sélectionner qu'une). D'autre part, l'elastic net a à plusieurs reprises obtenu de meilleures performances de prédiction que le LASSO (cf. Zou et Hastie (2005) sur des données issues de la biologie).

Comme d'autre méthodes apparentées la régression logistique pénalisée crée un prédicteur basé sur un modèle linéaire en assignant des poids β_i à chaque variable d'entrée $(1, \ldots, i, \ldots, p)$. Pratiquement, pour une pénalsiation elastic net le vecteur β est trouvé en résolvant le problème d'optimisation :

$$\hat{\beta} = \underset{\beta}{argmin}\ l(\beta) + \lambda \left(\frac{1-\alpha}{2} ||\beta||_2^2 + \alpha ||\beta||_1 \right)$$

où $l(\beta)$ est une fonction de perte à minimiser et les termes suivant correspondent aux normes 1 et 2 du vecteur de coefficient β par lesquelles est pénalisé le problème de minimisation. Le coefficient λ détermine l'importance de la pénalisation qui contraint les coefficients β_i a àller vers zéro. Le paramètre α détermine l'importance relative des deux normes dans la pénalisation, quand $\alpha = 1$ seule la pénalisation en norme 1 est utilisée (ce qui revient au LASSO), quand $\alpha = 0$ seule la pénalisation en norme 2 est utilisée (ce qui revient à la régression ridge, sans sélection de variable).

3.5 Résultats

Six types de modèles ont été considérés, ceux incluant une sélection de variables (SVM et régression logistique) et ceux sans (SVM uniquement), avec ou sans inclusion des attributs qualifiant la négation. L'interprétation de la sélection de variable a été faite à partir de modèles établi sur l'ensemble du corpus en utilisant les paramètres établis au cours de l'évaluation. Les résultats obtenus pour chacune des approches sont résumés dans la table 2.

Un des résultats les plus frappants est l'absence d'impact de la détection des environnement négatifs. Ce résultat est étonnant étant donné que la négation est présente dans 18,5% des critiques négatives contre 10,9% des critiques positives et qu'elle apparaît donc comme un paramètre prédicteur potentiellement pertinent. Par ailleurs, bien que la sélection de variables

	N. variables	Précision	Rappel	F-value
SVM	2829	88.18%	89.54%	88.84
SVM + nég.	3257	86.66 %	87.54%	86.77
Rég. logistique + sél. *elastic net*	1219	**88.78%**	**91.61%**	**90.16**
Rég. logistique + sél. *elastic net* + nég.	1028.7	87.77%	85.29%	86.49
SVM + sél. *elastic net*	1219	88.22%	90.32%	89.25
SVM + sél. *elastic net* + nég.	1028.7	86.92%	84.50%	85.66

TABLE 2 – Classification d'opinion : résultats

n'offre pas un réel gain de performance elle a l'avantage de réduire considérablement l'espace de variable, et donc de permettre une meilleure interprétation des modèles fournis.

Les résultats obtenus ne sont pas directement comparables à ceux rapportés dans la campagne DEFT'07 car nous n'avons ici pas considéré la catégorie "neutre" utilisée dans cette campagne. L'intégration de cette catégorie ferait baisser les performances relevées ici. On peut toutefois noter que nos performances se montrent supérieures à celles relevées par Pang *et al.* (2002) sur une tâche équivalente. Une explication à cette supériorité tient certainement d'une part à la généralité des modèles produits lors de l'apprentissage et à l'effet du pré-traitement des textes.

4 Conclusions

Le travail présenté ici a pour vocation de servir de base à l'exploration poussée du domaine de l'analyse de sentiment en français. Nous avons fourni des mesures de base pour une tâche de classification simple et montré l'intérêt d'appliquer des techniques de sélection de variable avant de procéder à un apprentissage automatique. Dans le futur, nous comptons nous appuyer sur ces premières expériences pour essayer d'améliorer les performances des modèles produits. À cet effet, nous prévoyons d'analyser plus en détail le cas de la négation et de son absence d'effet pour la classification par SVM. Plus généralement, un de nos objectifs est de mesurer l'intérêt d'ajouter de l'information sémantique dans les traits retenus, notamment en exploitant le caractère rhétorique de certains éléments linguistiques. Pour cela nous nous basons notamment sur les théories argumentatives du discours (Anscombre et Ducrot, 1983; Winterstein, 2010).

Une autre direction de recherche concerne l'interprétation des modèles produits. Si les modèles issus de l'utilisation de SVM sont généralement trop complexes pour être interprétés, le processus de sélection de variable s'offre mieux à l'interprétation puisqu'il met en avant les traits les plus pertinents pour la classification. Ainsi la sélection de variables présentée ici permet de confirmer l'importance de certaines catégories dans l'analyse d'opinion : les substantifs se retrouvent significativement sous-représentés dans les critiques négatives (45% des traits sélectionnés contre 54,1% avant sélection) au profit des adjectifs, verbes et adverbes (50,1% après sélection contre 44,6% avant). Par ailleurs, la sélection des connecteurs confirme les prédictions de certaines approches : la conjonction *mais* s'avère être un bon prédicteur de critique négative, alors que *et* est un prédicteur positif. En termes de stratégie argumentative (Winterstein, 2010), ces observations valident l'hypothèse qu'une critique positive va avoir tendance à présenter plusieurs arguments

positifs indépendants (reliés par *et*) alors qu'un seul argument négatif, même contre-balancé par un positif (avec le connecteur *mais*), suffira à produire une critique négative. Une autre approche dans cette perspective consiste à utiliser des techniques de *bootstrapping* qui permettent également d'évaluer l'importance des différents traits utilisés dans les processus d'apprentissage. Ces recherches sont actuellement en cours.

Références

ANSCOMBRE, J.-C. et DUCROT, O. (1983). *L'argumentation dans la langue*. Pierre Mardaga, Liège, Bruxelles.

DAS, S. et CHEN, M. (2001). Yahoo ! for amazon : Extracting market sentiment from stock message boards. *In Proceedings of the 8th Asia Pacific Finance Association Annual Conference (APFA 2001)*.

DENIS, P. et SAGOT, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation*, 46:721–746.

GHORBEL, H. et JACOT, D. (2011). Further experiments in sentiment analysis of french movie reviews. *In* MUGELLINI, E., SZCZEPANIAK, P. S., PETTENATI, M. C. et SOKHN, M., éditeurs : *Advances in Intelligent and Soft Computing*, volume 86, pages 19–28. Springer, Berlin.

GROUIN, C. et AL. (2007). Présentation de l'édition 2007 du défi fouille de textes (DEFT'07). *In Actes de l'atelier de clôture du 3ème Défi Fouille de Textes (DEFT'07)*, pages 1–8, Grenoble, France.

GUYON, I. et ELISSEEFF, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

JOACHIMS, T. (1999). Making large-scale svm learning practical. *In* SCHÖLKOPF, B., BURGES, C. J. C. B. et SMOLA, A. J., éditeurs : *Advances in Kernel Methods - Support Vector Learning*, pages 41–56. MIT Press.

PANG, B. et LEE, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

PANG, B., LEE, L. et VAITHYANATHAN, S. (2002). Thumbs up ? sentiment classification using machine learning techniques. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86. Association for Computational Linguistics.

VERNIER, M. (2011). *Analyse à granularité fine de la subjectivité*. Thèse de doctorat, Université de Nantes.

WILSON, T., WIEBE, J. et HOFFMANN, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354.

WINTERSTEIN, G. (2010). *La dimension probabiliste des marqueurs de discours. Nouvelles perspectives sur l'argumentation dans la langue*. Thèse de doctorat, Université Paris Diderot.

ZOU, H. et HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, Series B:301–320.