# Détection d'opinions dans des tweets

Caroline Collet*, Alexandre Pauchet*, Laurent Vercouter*, Khaled Khélif*

*LITIS-Avenue de l'Université - BP 8 76801 Saint-Étienne-du-Rouvray Cedex

**Résumé.** Twitter est à l'heure actuelle un des réseau sociaux les plus utilisé au monde et analyser les opinions qui y sont contenues permet de fournir de précieuses informations notamment aux entreprises commerciales. Dans cet article, nous décrivons une méthode permettant de déterminer l'opinion d'un tweet en détectant dans un premier temps sa subjectivité, puis sa polarité.

## 1 Introduction

De nos jours, les institutions étatiques tout comme les entreprises, s'appuient souvent sur l'opinion publique pour orienter leurs décisions stratégiques. L'analyse automatique d'opinions a ainsi connu une véritable envolée depuis l'apparition des réseaux sociaux tels que Twitter. Selon Pak et Paroubek (2010), une opinion peut être soit positive, négative, ou neutre, ce qui revient à un problème de classification en 3 classes. Un second point de vue, que nous adopterons, consiste à considérer qu'une opinion ne peut pas être neutre, seulement objective. Ainsi, le problème peut être décomposé en une évaluation de l'objectivité dans un premier temps, suivie dans le cas d'un texte subjectif d'une seconde étape de détection de la polarité.

L'objectif de ces travaux, dans le cadre d'un projet européen de tourisme, est de présenter une méthode pour détecter la subjectivité puis la polarité d'un tweet anglais ou français. Dans la suite de cet article, nous présentons section 2 un état de l'art sur la détection d'opinions dans des textes. La section 3 décrit notre méthode de détection de subjectivité et son évaluation sur des tweets. La section 4 est dédiée à la détection de polarité. Enfin, la section 5 résume et analyse les résultats obtenus et propose des pistes pour nos travaux futurs.

## 2 La détection d'opinions dans des textes

La détection de subjectivité et de polarité sont des tâches similaires pouvant être résolues par le même type de méthodes. Il en existe 3 à l'heure actuelle : les méthodes symboliques, statistiques et hybrides.

Les méthodes symboliques créent un ensemble de règles de décision servant à classer un texte dans une catégorie. Ces règles peuvent être éditées de manière manuelle (très coûteux) ou semi-manuelle par un expert humain et sont directement appliquées aux textes à évaluer. Les règles définies peuvent être plus générales (moins coûteuses) en utilisant par exemple des formules telles que des calculs de valences Baccianella et Sebastiani (2010); ¸Serban et Pécuchet (2012). Ces approches ne permettent néanmoins pas la détection d'opinions pour des structures linguistiques compliquées.

Les méthodes statistiques s'appuient sur un corpus de textes transformés en vecteurs de caractéristiques encodant les textes (comme le nombre d'occurrences) afin d'en extraire des propriétés par apprentissage. Le clustering ne nécessite pas de corpus annoté mais si les caractéristiques sont mauvaises, les classes obtenues ne correspondent pas aux classes désirées Hatzivassiloglou et McKeown (1997). Les méthodes telles que SVM nécessitent un corpus annoté mais sont plus efficaces. Pang et Lee (2008) a démontré que la seule présence des mots suffit à obtenir de bonnes performances.

Les méthodes hybrides mélangent apprentissage statistique et édition de règles (le plus souvent manuelles). La sélection de motifs séquentiels les plus fréquents Serrano et al. (2012) peut ainsi être classée dans la catégorie des méthodes hybrides puisqu'elle est constituée d'une partie d'extraction automatique de motifs les plus fréquents et d'une partie de sélection manuelle des motifs les plus pertinents pour un contexte donné. Ces méthodes semblent efficaces à condition d'effectuer les bonnes combinaisons.

Nous utiliserons SVM pour détecter la polarité puisque c'est une méthode très efficace et que Go et al. (2009) propose une méthode de constitution automatique de corpus. Pour détecter la subjectivité, en revanche, ne disposant pas de ressource d'annotation manuelle, nous utiliserons une méthode hybride sans corpus annoté : les motifs séquentiels fréquents.

# 3 Détection de subjectivité par motifs séquentiels fréquents

Pour détecter la subjectivité, les motifs séquentiels les plus fréquents sont d'abord extraits des tweets et, parmi eux, nous avons ensuite sélectionné manuellement ceux permettant de détecter la subjectivité. Si l'un des motifs au moins peut s'appliquer à un tweet, alors celui-ci est considéré comme subjectif.

Trois concepts de base sont utilisés dans cette méthode : les items, les itemsets et les séquences. Les items synthétisent les différentes informations sur un mot (le mot lui-même, son lemme - en français - ou son stem - en anglais - et sa catégorie grammaticale). Pour diminuer le bruit induit par les urls et les hashtags, nous les avons remplacé par le terme « URL » et le terme « HASH » et avons rajouté une annotation au termes subjectifs en utilisant Sentiwordnet comme référence. Les itemsets sont des ensembles d'items. Enfin les séquences sont un ensemble d'itemsets représentant un tweet. Il est ensuite nécessaire de transformer un corpus de tweets subjectifs en un ensemble de séquences. Nous avons constitué un corpus en ne récupérant que des tweets présentant un émoticône Go et al. (2009) (300 000 anglais et 300 000 français). Nous avons supprimé ceux ne possédant pas une entité nommée de type lieu touristique puisque le thème du projet est le tourisme. Nous obtenons 5 000 tweets pour l'anglais et 4 000 pour le français. Nous avons annoté nos propres corpus de test pour ne pas biaiser la validation en utilisant le corpus d'apprentissage puisque la méthode de Go et al. (2009) est naïve (700 pour le français et 800 pour l'anglais).

3 paramètres doivent être fixés pour affiner l'extraction de motifs : le nombre d'itemsets minimum et maximum à considérer pour générer les motifs (suffisamment grand pour obtenir des opinions et sans limite maximum), nous l'avons fixé à 3 pour ne pas couvrir les émotions ; le support, c'est à-dire la fréquence minimum d'apparition des motifs dans le corpus (suffisamment grand pour ne pas obtenir des règles trop précises mais suffisamment petit pour obtenir un nombre raisonnable de règles), nous l'avons fixé entre 100 et 200 pour obtenir 1000 motifs

environ ; et enfin le gap (nombre d'itemsets à ignorer lors de la génération des motifs. Un gap de 0 correspond à des itemsets situés côte-à-côte dans le texte).

Nous obtenons les meilleures performances avec un gap de 2, avec une précision moyenne de 64% et un rappel moyen de 62% pour l'anglais et une précision moyenne de 65% et un rappel moyen de 66% pour le français. Barbosa et Feng (2010) ont comparé quatre méthodes de détection de subjectivité pour l'anglais en utilisant le taux d'erreur (nombre de textes mal classés sur nombre total de textes) comme métrique. Les performances se situent entre 18 et 32. Dans notre cas, nous obtenons 34 pour le français et 33 pour l'anglais ce qui reste acceptable.

# 4 Détection de polarité

Pour détecter la polarité, nous sommes partis d'une méthode SVM avec un noyau non linéaire RBF afin de classer des textes. Nous avons choisi la présence des mots comme caractéristiques du vecteur puisque Pang et Lee (2008) la recommande. Chaque terme est stemmé (pour l'anglais) ou lemmatisé (pour le français) et associé à sa catégorie grammaticale. Les URL et les hastags qui pourraient bruiter l'apprentissage sont supprimés, ainsi que les déterminants. En suivant la méthode de Go et al. (2009) basée sur la polarité des émoticônes, nous sommes parvenus à créer un corpus annoté. Nous avons utilisé leur corpus de 1,6 millions de tweets pour l'anglais et avons constitué notre propre corpus français de 300 000 tweets. Pour valider la méthode, nous avons voulu éviter une validation croisée étant donnée que la méthode d'annotation est naïve. Nous avons ainsi annoté nos propres corpus de test (700 tweets français et 800 anglais). Il a ensuite été nécessaire de sélectionner les termes à conserver dans l'index. Nous avons commencé par conserver les 1 000 plus fréquents pour l'anglais et les 10 000 plus fréquents pour le français. En effectuant un rapide test, nous avons pu nous apercevoir qu'en augmentant ou en diminuant ce nombre, nous perdions en précision moyenne et en rappel moyen. Ensuite, en observant les tweets classés par le système, nous avons pu constater que la négation n'est généralement pas prise en compte. Nous proposons une méthode alternative aux n-grams, moins efficaces que la seule présence des motsPang et Vaithyanathan (2002), basée sur la position absolue des mots dans la phrase. Nous choisissons d'inscrire non plus la présence des mots dans l'index mais la position de chaque terme pondéré par la taille du tweet. Par exemple la phrase « it is time now » aura 0.25 pour it (position 1, tweet de taille 4), 0.5 pour is (position 2), 0.75 pour time (position 3) et 1 pour now (position 4). L'inconvénient est qu'un même mot peut apparaître plusieurs fois dans la phrase. Il devra donc apparaître autant de fois dans l'index.

En utilisant la seule présence des mots, nous obtenons 74% de précision moyenne et 73% de rappel moyen pour l'anglais et 62% de précision moyenne et 61% de rappel moyen pour le français. Avec la position absolue, les performances diminuent à 60% pour la précision et le rappel en anglais. En revanche, les performances françaises sont équivalents : 62.6% de précision et 57% de rappel. Compte-tenu des résultats, il est probable que l'information de position absolue soit trop précise et qu'il faille plutôt considérer une position relative. En comparant avec Go et al. (2009), ceux-ci ont obtenu 82% d'accuracy sur leur corpus de test anglais, dans notre cas, nous atteignons tout de même 73% sur un corpus de test différent.

# 5 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode permettant de détecter la subjectivité et la polarité d'un tweet. Dans le cas de la détection de subjectivité les performances sont correctes mais inférieures à l'état de l'art puisque les motifs retournés présentent un bruit important dû aux émoticônes et à la ponctuation. Il sera donc nécessaire de les supprimer du corpus. De plus, les mots subjectifs ont tendance à disparaître dans les motifs remplacés par leur catégorie grammaticale, ce qui ne permet pas une sélection manuelle efficace. Il sera donc nécessaire de mettre en place une méthode pour que ces termes apparaissent dans les motifs. Enfin, dans le cas de la détection de polarité, nous avons obtenu des performances raisonnables bien que légèrement inférieures à ceux de la littérature, ce qui peut s'expliquer par une difficulté à sélectionner les paramètres optimaux du SVM. Une piste d'amélioration consisterait à supprimer de l'index les termes présents de manière équivalente dans les deux classes et d'ajouter les termes subjectifs non conservés car moins fréquents, en utilisant Sentiwordnet par exemple.

# Références

Baccianella, E. et Sebastiani (2010). Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion minings. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Barbosa, L. et J. Feng (2010). Robust sentiment detection on twitter from biased and noisy data. In J. Huang, Chu-Ren et Dan (Eds.), *COLING (Posters)*, pp. 36–44. Chinese Information Processing Society of China.

¸Serban, Pauchet, R. et Pécuchet (7 pages, 2012). Semantic propagation on contextonyms using sentiwordnet. In *Workshop Affects, Compagnons Artificiels et Interaction*, Grenoble.

Go, A., H. Bhayani, Richa, et Lei (2009). Twitter sentiment classification using distant supervision. *Processing*.

Hatzivassiloglou et McKeown (1997). Predicting the semantic orientation of adjectives.

Pak et Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion mining.

Pang et Lee (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*.

Pang, L. et Vaithyanathan (2002). Thumbs up ? sentiment classification using machine learning techniques. *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*.

Serrano, C., G. Brunessau, et Bouzid (2012). Combinaison d'approches pour l'extraction automatique d'événements. *TALN'2012*.

# Summary

Twitter is nowadays one of the most used social network in the world and analysing the opinions inside gives precious informations to business society. In this article we describe a method that enables to determine opinions in tweets by detecting first the subjectivity and then the polarity.