Synergies Pologne n° 13 - 2016 p. 113-121

GERFLINT
ISSN 1774-7988
ISSN en ligne : 2261-3455

La polarité des avis des internautes[1] : repérage automatique

Alicja Hajok
Université Pédagogique de Cracovie, Pologne
alicjahajok@gmail.com
Luis Meneses Lerin
Université d'Artois, France
GRAMMATICA (EA4521)
luis_meneses_lerin@yahoo.fr

Résumé

Cet article expose les résultats d'une étude menée dans le cadre du projet POLONIUM. Ce projet cherche à proposer une méthode du type « hybride » (statistique et linguistique) pour le traitement automatique des avis, sentiments ou opinions des internautes.

Mots-clés : polarité, domaine, contexte, corpus, étiquetage sémantique, opinions

The polarity: Diverging Opinion among Internauts

Abstract

In this paper, we present the results of a research conclucted as part of the POLONIUM project. The project aims to propose a "hybrid" method for the processing of opinions and feelings which takes into account both statistical and linguistic parameters.

Keyword : polarity, linguistics, corpus, context, field, opinion

1. Problématique, objectifs et résultats

Les nouvelles technologies ont influencé le mode de fonctionnement de notre société. Les individus communiquent et échangent avec d'autres individus ou avec la société entière sur les réseaux sociaux, les commentaires, les blogs, etc. De nos jours, l'internaute n'est plus qu'un simple *consommateur* passif d'information, mais il est devenu un acteur actif qui « génère » de l'information. Cela a été possible grâce à la démocratisation des dispositifs actuels (ordinateurs, tablettes, portables) et à l'évolution d'outils informatiques capables d'être manipulés par pratiquement n'importe quel individu. C'est ainsi que la société utilise Internet dans ses activités

quotidiennes pour transmettre des quantités extraordinaires d'information en abordant n'importe quel sujet. Internet, et plus précisément les réseaux sociaux, permet aux différentes communautés d'internautes non seulement de traiter des sujets très précis, mais également d'organiser des mouvements sociaux ou des révoltes (*printemps arabe, Soy 123, Je suis Charlie*, etc.). D'où l'importance de suivre et d'observer ce qui est « dit » dans ces moyens de communication afin de connaître le « sentiment » ou les « opinions » d'une communauté ou d'un peuple en particulier.

L'un des défis actuels concerne le traitement automatique des commentaires des internautes dans les réseaux sociaux, y compris, les commentaires postés sur internet et les blogs. Ce type de texte possède certaines caractéristiques qui se démarquent, par exemple, des textes journalistiques. Il s'agit d'un texte hybride qui mêle le langage écrit et oral. Les méthodes actuelles de « fouille de textes » s'avèrent pertinentes pour les analyses des textes journalistiques. Ce qui n'est pas le cas avec des messages postés par les internautes. L'approche traditionnelle de la fouille de textes se sert de la fréquence d'apparition des mots pour identifier et classer des catégories de documents de texte (Sebastiani, 2002). Cependant cette méthode s'avère peu appropriée lorsqu'il s'agit des messages (avis) de la part des internautes dans des réseaux sociaux. D'où l'importance de proposer de nouvelles approches afin de mesurer la similitude entre ces types de texte.

2. Le corpus de travail

Actuellement, les systèmes qui classifient les commentaires des internautes (positifs, négatifs, neutres) sont basés sur des systèmes purement statistiques. Ces systèmes ne tiennent pas compte de la nature des « avis », à savoir de leur nature « linguistique ». Souvent, les systèmes informatiques dans le domaine du Traitement Automatique des Langues oublient que l'étude de mots isolés ne s'avère pas pertinente et que, si l'on souhaite extraire ou classer des informations, il est important de tenir compte du « cotexte » ou du « domaine ». Nous cherchons à proposer un traitement linguistique à partir d'un corpus d'entraînement qui tiendra compte du *cotexte* et du *domaine* afin d'extraire et de classer les données en question. Nous proposerons, d'une part, une application informatique qui utilisera une méthode « hybride » avec une *composante statistique* et une *composante linguistique* pour ensuite effectuer une expertise linguistique et évaluer les résultats obtenus. Cela nous permettra de proposer une étude de la micro-distribution des termes et l'extraction automatique des univers sémantiques par des biais : (i) d'extraction d'un univers thématique – gravitant autour d'un mot clé, (ii) du recensement des cooccurrents et des séquences d'items et (iii) de l'analyse de la couverture phraséologie et collocationnelle (Grossmann & Tutin, 2003, Mejri, 1997).

Le corpus analysé contient des commentaires des tweets[2] concernant le « changement climatique ». Il a été constitué par le centre de recherche LABTL dirigé par Luis Villaseñor-Pineda[3]. Ce corpus doit être considéré comme un échantillon de données, car il ne peut pas être considéré comme représentatif compte tenu de la taille gigantesque de tweets que nous pouvons trouver concernant le sujet retenu. Du point de vue quantitatif, l'échantillon de données contient 300 commentaires des utilisateurs de Twiter concernant le changement climatique. Ces commentaires ont été annotés par l'équipe LABTL en proposant trois types d'étiquettes : *polarité positive*, *polarité négative* et *polarité neutre*. Chacun des commentaires (tweets) a été annoté selon la polarité du commentaire par rapport au *changement climatique*. L'annotation a été réalisée de manière automatique et a utilisé une base de données d'unités lexicales présentant un certain degré de subjectivité. Grâce à des algorithmes capables de mesurer la similitude entre les unités lexicales contenues dans la base de données et celles présentes dans les tweets, il a été possible de proposer un étiquetage pour évaluer la polarité de chacun des tweets.

Voici un exemple des données contenues dans l'échantillon qui fonctionne comme corpus afin d'évaluer la pertinence des annotations proposées :

1. *#smurfitkappa renforce son engagement en faveur de la lutte contre le changement climatique http://t.co/2S7RIhWHbU ', positive*
2. *La ville, écosystème du XXIème siècle http://t.co/yQQPWrMM3T ', neutre*
3. *@Kyoht putain changement climatique. ', négative*

3. À propos de la polarité et de l'E-polarité

Le Grand Robert définit la polarité comme : « Propriété qu'a l'aimant ou une aiguille aimantée de se diriger vers les pôles du monde », il s'agit donc d'un « état d'un système dont deux points quelconques présentent des caractéristiques différentes (opposées ou distinctes) ». En linguistique, les travaux abordent souvent la question de la *polarité* en se limitant à la *polarité négative* (Larrivée, 1995 & 2004) et on la lie à d'autres phénomènes comme la négation, la quantification et la présupposition. Par exemple, C. Muller précise (1991 : 69) repris par S. Palma, Présentation à Langages n°162, *Polarité, négation et scalarité*, 2006 : 3 qu'«*il s'agit* de *phénomènes d'influence* du *contexte* sur la *possibilité d'occurrence* ou le *sens d'expressions qui y sont sensibles*. *Pratiquement*, on *peut définir* les *termes* à *polarité négative*, et les *contextes* à *polarité négative*». Giannakidou 2008 constate que les travaux abordent la *polarité négative* à travers des débats qui utilisent comme facteur déterminant la véridicité (idem) : (i) la « polarité négative » est

aussi abordée uniquement du point de vue lexical et syntaxique ; (ii) elle fait partie d'une catégorie d'expression que l'on retrouve pratiquement dans toutes les langues : la « négation ». D'autres travaux ont abordé le sujet de la négation (Horn 1985 et, pour le français, Larrivée 2004) qui peut être marquée linguistiquement à l'aide de la morphologie (préfixes), de la syntaxe (adverbes, coordonnants, déterminants et pronoms, modification des syntagmes et du groupe verbal), de la sémantique (unités lexicales, collocations et locutions) et de la pragmatique (prise en compte des éléments intralinguistiques et extralinguistiques).

Cependant la nature de la *polarité positive* et *neutre* reste peu décrite et étudiée (Larrivée, 2012). On retrouve encore d'autres termes peu étudiés comme la *polarité opposée* qui est définie comme « le fait d'admettre deux propriétés antonymes (exprimées par des adjectifs 'à polarité opposée'), comme une caractéristique de la comparaison de déviation *Pierre est aussi aimable que Paul est désagréable* (Fuchs, 2014 : 85) et encore la polarité congruente qui renvoie aux « comparaisons à déviation [qui] peuvent aussi jouer sur deux propriétés co-orientées (c'est-à-dire de polarité congruente) prédiquées de deux entités ou d'une même entité dédoublée […] : *…le lecteur aura effectivement des idées en plus et des embarras en moins* (Anscombre in Fuchs, 2014 : 85).

En sortant de ces constats, nous dirons que l'E.-polarité est un état d'un système dont deux points quelconques présentent des caractéristiques différentes (opposées ou distinctes). On dégage les marques de la polarité (*c'est bien ≠ ce n'est pas bien*) au sein du tweet analysé qui permettent de le classer comme positif ou négatif. Cependant, ces marqueurs ne couvrent qu'une partie des tweets ce qui ne répond pas aux besoins d'exhaustivité d'analyse. Ainsi il est nécessaire de proposer des analyses approfondies du contexte et du cotexte des tweets :

- le cotexte proche de l'unité linguistique analysée, autrement dit ses occurrences à gauche et à droite, par exemple *c'est mal, mais tant pis* et *c'est pas mal*.
- le contexte linguistique plus large dans lequel est publié le tweet en question, donc (i) un texte (un article, un événement) sur lequel porte le tweet et (ii) tout un enchaînement de tweets et de retweets ce qui demande un suivi qui prend en compte aussi les éléments cataphoriques et anaphoriques.
- le contexte qui prend en compte nos connaissances extralinguistiques.

En bref « les choses que vous dites n'existent que dans le contexte d'autres communications et […] on ne peut pas les regarder de manière isolée, comme si elles étaient des publications uniques et singulières » (Paveau, 2006). L'analyse du contexte et du cotexte permettra de dégager le tissage qui existe entre tous

ces éléments. Nous pouvons même parler de la mémoire discursive des tweets. La *mémoire discursive* est définie comme l'«ensemble des savoirs consciemment partagés par les interlocuteurs». Elle « n'est pas tant alimentée en permanence par des événements de la situation extralinguistique que par les énoncés portant sur ces événements et constituant eux-mêmes des événements ». En effet, tout énoncé, aussi bref qu'il soit, a toujours besoin d'un co(n)texte (Adam, 2011 : 38-42).

Enfin, si un tweet ne présente pas de polarité positive ou négative, nous parlerons de la polarité neutre. Nous l'illustrerons dans la partie suivante.

4. Quelques critères linguistiques pour mesurer la polarité

A l'état actuel de nos études, nous avons dégagé trois étapes d'analyse linguistique des tweets qui consiste à viser (i) les mots clés, (ii) les unités de la langue générale à polarité (négative, positive, neutre) et (iii) les émotions - positives ou négatives (les marqueurs d'émotion).

(i) Viser les mots clés

Nous avons décidé de travailler sur les tweets qui renvoient à des sujets concernant le domaine de l'écologie et facilite la première étape d'analyse. Or, la langue de spécialité se maîtrise plus facilement que la langue générale. Ensuite, les termes spécialisés retenus et ensuite employés par le grand public sont relativement restreints. Dans la majorité des cas, seulement les termes les plus courants sont repris du texte commenté et ensuite sont employés dans les tweets et retweets. Alors, il s'agit des mots spécialisés à un domaine précis qui sont statistiquement fréquents et ils sont souvent précédés du croisillon autrement dit du hashtag (#) (mot-dièse ou mot-clic) dont l'objectif est de centraliser les messages autour d'un terme bien précis (ex. 7 #climat).

Nos connaissances extralinguistiques permettent de prédire, de façon très subjective, la polarité du terme. Prenons deux exemples :

Les unités spécialisées de la *polarité positive :*

4. *488389360929083392,' Libellule : la voiture écologique aux 1000 km d'autonomie ! http://t.co/TXAwg0cTlG ', positive*[4]

Les unités spécialisées de la *polarité négative :*

5. *487735557334659072,' Réchauffement climatique et d'autres problème du changement climatique affecte tout le monde à certains, http://t.co/AeJziCYyCA ', négative*

Dans (4), nous avons visé l'adjectif *écologique* qui modifie le substantif *voiture*. Nos connaissances extralinguistiques permettent de paraphraser cette structure comme : « *une voiture écologique est une voiture qui ne pollue pas donc il s'agit d'une voiture qui respecte l'environnement*, etc. D'ailleurs, nos pressentiments sont confirmés par le Grand Robert qui propose une définition suivante : *écologique adj. – (…)* Cour. *Qui respecte les équilibres écologiques naturels*. La polarité positive est en plus renforcée par l'unité *1000 km d'autonomie*. Nos connaissances extralinguistiques ou plutôt celles du « chauffeur ordinaire » permettent de voir dans *1000km* une très bonne autonomie.

Dans (5), nous retenons deux unités *réchauffement climatique* et par conséquent *le changement climatique*. Nous lisons dans Wikipedia que le réchauffement climatique: « c'est un phénomène d'augmentation des températures sur la plus grande partie des océans et de l'atmosphère terrestre (…). Néanmoins l'impact économique, sociologique, environnemental, voire géopolitique, de ces projections *est globalement négatif* à moyen et long terme »[5]. Cette vision négative du terme est encore reprise par l'unité linguistique *problème* qui est vue toujours par la prise de la négation : « Difficulté qu'il faut résoudre pour obtenir un résultat ; situation instable ou dangereuse exigeant une décision. Ennui » (Le Grand Robert). Du point de vue linguistique c'est l'unité « problème » qui a créé l'interprétation négative dans ces exemples.

Il est donc indispensable de regrouper les termes spécialisés selon la polarité, positive, négative ou neutre, qu'ils véhiculent. Cependant, cette classification reste très subjective. Comparons :

6. *487702882138198016,' @Poupinouik @heavens66 tu as raison, puis si je me prends un vent, pas grave, les éoliennes, c tendance #MerciSégolène ', positive*

7. *487894866136166400,' #climat Les éoliennes, arme de torture massive pour les êtres vivants. http://t.co/riK55KcK8J … http://t.co/GTKnStcoKU ', négative*

Dans (6) et (7), le terme *éoliennes* reste ambiguë. Pour les uns, c'est un symbole d'énergie propre, du développement durable (6). Cependant, il serait intéressant de voir un contexte plus large, car *pas grave* et *#MerciSégolène* font ressentir l'ironie. D'ailleurs, la question d'ironie demande des études plus approfondies.

Pour les autres, c'est un symbole de nuisance sonore, des infrasons dangereux pour la santé (7). La désambiguïsation de ce terme peut se faire seulement en tenant compte du contexte. Ce sont les termes de la langue générale comme *tendance* et *arme de torture massive* qui penchent la polarité d'un côté à l'autre.

(ii) Viser les unités de la langue générale à polarité (négative, positive, neutre)

Les prédicats et les actualisateurs[6] (Hajok & Mejri, 2011) issus de la langue générale (aussi bien que la langue soutenue et la langue familière) expriment la polarité positive ou négative.

Les unités de la langue générale de la polarité positive :

8. *487549421090267136,' #smurfitkappa renforce {VS} son engagement {Npred} en faveur de[7] la lutte contre {Vpred} le changement climatique http://t.co/2S7RIhWHbU ', positive*

9. *487889930740170753,' Louis Giscard d'Estaing répertorié dans les 18 maires de France favorables aux voitures écologiques par son action http://t.co/J61mqZmxwC ', positive*

Les unités de la langue générale de la polarité négative :

10. *487735557334659072,' Réchauffement climatique et d'autres problème {Npred} du changement climatique affecte tout le monde à certains, http://t.co/AeJziCYyCA ', négative*

11. *487643906377805824,' @bertin85 Des éoliennes proches d'habitation sont une nuisance {Npred} ; de plus, elles sont des tueuses{Npred} pour nos oiseaux (via LPO)@Valeurs @coolise ', négative*

Les tweets étiquetés comme « neutres » sont dépourvus de ces marqueurs linguistiques (positifs ou négatifs) par exemple les textes du type informatif :

12. *488357032181788673,' 07<07<14 - Conférence citoyenne sur le développement durable http://t.co/NQ11WX2MWy ', neutre*

13. *487983506505887744,' Mode d'emploi du réseau social Humanite-Biodiversite.fr http://t.co/FgwbJkt9Z4 ', neutre*

(iii) Viser les émotions - positives ou négatives (les marqueurs d'émotion)

Quelques formes expressives retenues dans les tweets :

a) les exclamations et onomatopées : *#grr, #beurk, #arfff, #pfff*

14. *488270492357623808,' @cdion @OfficielML # bravo #demain check this http://t.co/QSCPgHL0Aj une invention qui va bouleverser le marché des panneaux solaires. Bon vent ', positive*

b) les images réalisées à l'aide du code ASCII :

15. *488392451737284609,' @PaulMaxit @SophieCamard @pcanfin @MicheleRivasi les oiseaux et les éoliennes LOL les chats sont biens plus violents ! ', positive*

c) les formes lexicales explicites et précédées par #: #colère, #joie, #scandalisée :

16. *488669757542248448,' Un député EELV justifie les attaques sur les synagogues ! C'est un #scandale d'en arriver là... http://t.co/7YqCK-MuKzQ] ', négative*

d) les formes lexicales figées :

17. *487387097222098944,'Développement durable ma gueule, {SF_ Interjection_pred}, négative*

Conclusion

Nous venons d'exposer les premiers résultats d'une étude menée dans le cadre du projet POLONIUM. L'objectif de ce projet est de proposer une méthodologie permettant l'étiquetage de façon automatique des relations existantes entre les unités linguistiques de tweets. En bref, il s'agit d'identifier et d'interpréter la polarité, positive, négative ou neutre, de n'importe quel tweet.

Bibliographie

Adam, J.-M. 2011. *La linguistique textuelle*, 3ième édition, Paris : Armand Colin.

Anscombre, J.-C., Mejri S. (éds.) 2011. *Le figement linguistique : la parole entravée*. Paris : Honoré Champion.

Fuchs, C. 2014. *La comparaison et son expression en français*, Editions OPHRYS, 208.

Giannakidou, A. 2008. « Negative and positive polarity items: Variation, licensing, and compositionality ». In: Maienborn, C., Klaus von Heusinger, and Portner P. (eds.*Semantics: An International Handbook of Natural Language Meaning*. Berlin: Mouton de Gruyter. *semarch. linguistics.fas.nyu.edu/Archive/.../handbookpaper.pdf*

Goes, J, Pitar, M. 2013. *La négation : Etudes linguistiques, pragmatiques et didactiques*. Arras : Artois Presses Université.

Grossmann, F., Tutin, A. (eds.) 2003. *Les collocations : analyse et traitement*, Travaux et recherches en linguistique appliquée, Amsterdam : de Werelt.

Hajok, A., Mejri, S. 2011, *Neophilologica*. Vol. 23: *Le figement linguistique et les trois fonctions primaires (prédicats, arguments, actualisateurs) et autres études*, Université de Silésie, Pologne.

Horn, L.R. 1985. « Metalinguistic negation and pragmatic ambiguity », *Language* 61, 121-174.

Larrivée, P. 1995. Négation et polarité négative : gradient de valeurs extrêmes. In : *Cahiers de l'Institut de linguistique de Louvain*, 21.

Larrivée, P. 2004. *L'Association négative : depuis la syntaxe jusqu'à l'interprétation*. Genève : Droz.

Larrivée, P. 2012. « Positive Polarity, Negation, Activated Propositions », *Linguistics* 50, 4, 869-900.

Mejri, S. 1997. *Le figement lexical : descriptions linguistiques et structuration sémantique*, Publications de la Faculté des lettres de la Manouba, série linguistique, volume X, Tunis.

Meneses-Lerín, L. 2014. « Mot et emplois : la problématique de l'unité croisée ». In : 9èmes Journées Scientifiques du réseau Lexicologie, Terminologie, Traduction, Laboratoire. *L'unité*

en Sciences du langage, M. Van Campenhoudt, I. Sfar et S. Mejri (dirs.), Publications de L'actualité Scientifique, AUF. p. 87-102

Muller, C. 1991. *La négation en français*. Genève : Droz.

Palma, S. 2006. *Langages* : « Polarité, négation et scalarité », 40e année, n°162.

Paveau, M.-A. 2006. « L'intégrité des corpus natifs en ligne. Une écologie postdualiste pour la théorie du discours ». *Les cahiers de praxématique*, Montpellier : Presses universitaires de la Méditerranée, 2006, 2015, *Corpus sensibles*, p.65-90. <hal-01185710>

Sebastiani, F. 2002. « Machine Learning in Automated Text Categorization ». ACM Computing Surveys, Vol. 34, N° 1 *nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf* [Consulté le 02 mars 2016].

Notes

1. Cet article a été rédigé grâce au soutien accordé par le programme PHC 2016 POLONIUM de l'Ambassade de France en Pologne et Le Ministère des affaires étrangères et du développement international en Pologne. Ce projet est réalisé sous la direction de Alicja Hajok de l'Université Pédagogique de Cracovie, Pologne et de Luis Meneses Lerin de l'Université d'Artois, France. Projet N° 35371SF.

2. Un tweet est un texte hybride limité à 140 caractères qui mêle le langage écrit et le langage oral. Il existe des tweets dont l'écriture est linéaire, d'autres combinent librement des mots, des liens URL et des hyperliens, des symboles, des formes iconiques, des émoticônes et des images réalisés à l'aide des lettres et des caractères spéciaux contenus dans le code ASCII et des signes spéciaux qui organisent l'information (#, @, /, et d'autres) ce qui facilite l'exploitation de données et l'extraction de données

3. LabTL, Instituto Nacional de Astrofísica, Óptica y Electrónica, Computer Science Department, Mexique

4. Nous n'apportons aucune modification ni orthographique ni grammaticale aux tweets cités.

5. https://fr.wikipedia.org/wiki/R%C3%A9chauffement_climatique. [Consulté le 02 mars 2016].

6. Une unité linguistique peut jouer dans une phrase une de trois fonctions primaires: prédicative, argumentale ou actualisatrice.

7. De part de son origine *en faveur de* présente un sous-codage positif.