Combiner analyses textométriques, apprentissage supervisé et représentation vectorielle pour l'analyse de la subjectivité

Egle Eensoo, Damien Nouvel, Amélie Martin, Mathieu Valette

Combiner analyses textométriques, apprentissage supervisé et représentation vectorielle pour l'analyse de la subjectivité

Egle Eensoo[1] Damien Nouvel[1] Amélie Martin[1,2] Mathieu Valette[1]
(1) ERTIM, INALCO, 2 rue de Lille 75007 Paris
(2) SNCF Innovation et Recherche, 40 avenue des Terroirs de France, 75012 Paris
egle.eensoo@inalco.fr, damien.nouvel@inalco.fr, amelie.martin2@sncf.fr,
mathieu.valette@inalco.fr

Résumé. Cet article présente le bilan de notre participation au Défi Fouille de Textes (DEFT 2015) pour les tâches 1 et 2. Il s'agit de classer un corpus de tweets selon leur polarité (tâche 1) et détecter les classes génériques (tâche 2.1) et spécifiques (tâche 2.2) de ces derniers. Nous avons implémenté deux systèmes pour ce défi. La première méthode repose sur la sélection dans le corpus d'entraînement d'un ensemble de descripteurs sémantiquement motivés pour chaque tâche à partir d'une analyse textométrique, qui sont ensuite injectés dans un algorithme d'apprentissage automatique supervisé, permettant le calcul de modèles sur ce même corpus. La seconde méthode s'appuie sur une représentation vectorielle des mots apprise par utilisation de l'outil word2vec sur un corpus hétérogène et volumineux, cette représentation étant ensuite utilisée pour réaliser un apprentissage automatique supervisé, pour chaque tâche, sur les corpus de développement. Un troisième système a été réalisé par combinaison des deux précédents à l'aide d'heuristiques simples. Les résultats obtenus sur les corpus de tests montrent que chaque méthodologie a ses avantages et que leur combinaison peut donner de très bonnes performances.

Abstract.

Combining Textometric Analysis, Machine Learning and Vector Space Representation for Subjectivity Analysis. This paper reports the results of our participation in Evaluation Campaign of Text Mining (DEFT 2015) for tasks 1 and 2. The aim is to classify tweets according to their polarity (Task 1) and detect the generic (task 2.1) and specific classes (task 2.2) thereof. We implemented two systems for this challenge. The first method is based on the selection in the training corpus of a set of semantically motivated descriptors for each task from a textometric analysis, which are then injected into a supervised machine learning algorithm, allowing the development of models on the same corpus. The second method is based on a vector representation of words learned by using the tool of word2vec leveraging heterogeneous and large corpora. This representation is then used to perform automatic supervised learning, for each task, on the development corpus. A third system was designed by combination of both, using simple heuristics. The results obtained on the test corpora show that each methodology has its advantages and that their combination can achieve very high performance.

Mots-clés : analyse de la subjectivité, textométrie, word2vec, classification automatique, linguistique de corpus.

Keywords : subjectivity analysis, textometry, word2vec, machine learning, corpus linguistics.

1 Introduction

1.1 Campagne DEFT 2015

La fouille de données subjectives (sentiments, opinions, émotions) est depuis plusieurs années maintenant un domaine très dynamique de la fouille de textes, aussi bien dans le domaine académique que dans l'industrie. Sommairement, on observe quatre tendances en termes de positionnement épistémologique : méthodes par apprentissage (Pang *et al.*, 2002), méthodes symboliques d'inspiration cognitiviste (vocabulaire des émotions, etc. (Ghorbel & Jacot, 2011 ; Maurel & Dini, 2009), méthodes symboliques d'inspiration pragmatique ou analyse du discours (Vernier *et al.*, 2009a,b), méthodes hybrides combinant certaines de ces approches (Turney, 2002 ; Yi *et al.*, 2003 ; Yu & Hatzivassiloglou, 2003).

La campagne d'évaluation DEFT 2015 propose des tâches de détection de subjectivité (opinions, sentiments et émotions)

sur les tweets en français portant sur la thématique de changement climatique. Nous avons participé aux trois tâches suivantes :

— **Tâche 1** : La première tâche vise à classer les tweets selon une grille macroscopique de polarité : positif, négatif, neutre (ou mixte).
— **Tâche 2.1** : Cette tâche consiste à identifier le type de subjectivité (ou d'objectivité). Les classes proposées sont les suivantes : information (tweet objectif), opinion (l'expression intellective et réfléchie), sentiment (l'expression intellective-affective) et émotion (l'expression purement affective).
— **Tâche 2.2** : Dans cette tâche, l'objectif est d'identifier une classe fine correspondant à trois catégories subjectives (opinion, sentiment, émotion). DEFT nous propose 18 classes fines.

1.2 Travaux précédents et positionnement méthodologique

Notre participation au DEFT 2015 a été motivée par nos travaux antérieurs (Eensoo & Valette, 2012, 2014b,a) portant sur la détection d'opinions et l'analyse des sentiments sur divers corpus issus essentiellement du Web 2 (forums de discussions, commentaires d'internautes des articles de presse). Ainsi, nous avons pu élaborer une méthodologie qui s'inspire de la sémantique textuelle (Rastier, 2001) pour identifier des critères linguistiques pertinents pour une classification sémantique des textes subjectifs. Cette méthodologie s'appuie sur une analyse différentielle du corpus par des méthodes de textométrie comme le calcul de spécificités (Lafon, 1980), de collocations (n-grammes) et des cooccurrences (Lafon, 2981). Ces travaux se démarquent des approches traditionnelles fondées sur la recherche de marqueurs axiologiques explicites par l'utilisation de critères qui ne sont pas considérés d'ordinaire comme prioritaires pour la détection de l'information subjective. Ils relèvent des représentations des acteurs (composante dialogique), des structures argumentatives et narratives des textes (composante dialectique) et des thèmes instanciés (composante thématique). Nous avons pour objectif de proposer une méthodologie mixte alliant l'analyse du linguiste qui, en expertisant le corpus en extrait les éléments linguistiquement pertinents pour l'expression de la subjectivité et les méthodes statistiques qui automatisent l'analyse du corpus et rendent les résultats reproductibles. Les deux verrous scientifiques auxquels nous confrontons notre méthodologie en participant au défi DEFT 2015 ont trait au genre textuel du tweet d'une part, et à l'annotation fine d'autre part.

1. *Textualité et forme brève.* Notre méthodologie repose en effet sur une analyse sémantique de la textualité (cohésion textuelle, marqueurs structuraux etc.). Le tweet, forme brève réputée parataxique, est intrinsèquement pauvre en marqueurs de textualité et peut s'apparenter à un ensemble de mots-clés faiblement articulés textuellement. Conséquence probable de cette pauvreté textuelle, le tweet est hyperlexicalisé, comme en atteste l'innovation du mot-dièse (hashtag) qui promeut les mots du texte et parfois même des syntagmes complexes, voir des phrases au rang de mots-clés ou de candidats mots-clés. Notre méthodologie est conçue pour évaluer la capacité classificatoire des différents marqueurs sémantiques en particulier non thématiques et non axiologiques en privilégiant les éléments de structuration des textes et de positionnements énonciatifs (Eensoo & Valette, 2015). Elle serait donc peu adaptée à un genre textuel court donnant *a priori* le primat aux lexèmes porteurs de signification référentielle.
2. *L'annotation fine.* Il apparaît que l'annotation fine du corpus est en fait une annotation lexicale. C'est peut-être un corollaire du premier verrou scientifique : le guide d'annotation avec lequel le corpus semble avoir été produit [1] apparaît orienté vers une catégorisation très lexicale des tweets. Autrement dit, c'est davantage les significations des unités lexicales de chaque tweet, prises isolément, qui font l'objet d'annotation que le sens du tweet pris dans son ensemble. Au fond, on est ici confronté à une imprécision méthodologique. L'annotation fine ne signifie pas que les émotions vont être annotées avec finesse mais en fonction des seuls mots du texte, considérés comme des mots-clés indexant des émotions. En définitive, annotation fine signifie à grain fin (le grain étant celui du mot). À titre d'exemple, on peut reprendre celui du guide d'annotation « L'amour et la fidelité sont des espéces en voie de disparition » (orthographe respectée). Ce tweet est annoté AMOUR mais il est manifeste que l'émotion exprimée ici n'est pas l'amour, elle est vraisemblablement déceptive (pessimisme, consternation, résignation) mais pourrait également être, dans une perspective nihiliste, l'espoir, la joie (« enfin, nous voilà débarrassés de l'amour et de la fidélité »). En bref, on ne peut guère statuer sur l'émotion exprimée. De la même façon, un tweet tel que « Moi aussi j'aime l'entreprise, celle sans patron, sans actionnaire et qui produit des biens et services durables socialement et écologiques ! » (extrait du corpus) peut-il sérieusement être annoté comme porteur de l'émotion AMOUR ? la seule présence du verbe aimer ne permet pas, selon nous, d'en juger. On pourrait même argumenter que l'amour de l'entreprise exprimé ici l'est par contraste avec un désamour tout aussi explicite et même peut-être plus saillant

1. https ://deft.limsi.fr/2015/guideAnnotation.fr.php ?lang=fr

envers d'autres acteurs du tweets : le patron, les actionnaires. Nous développerons cette analyse critique dans le paragraphe 2.1.1.

2 Méthodologie de la détection de subjectivité

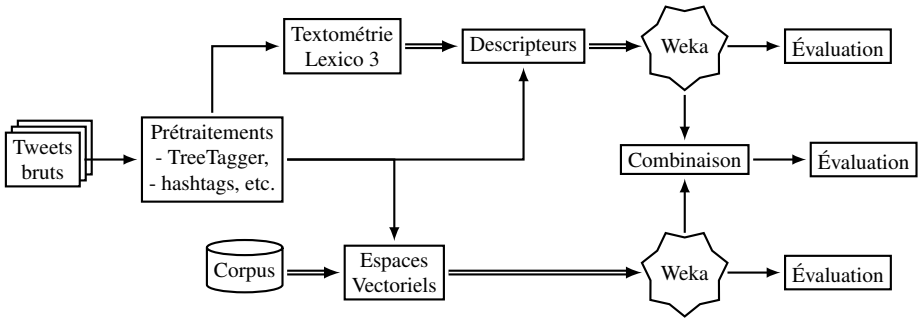

FIGURE 1 – Fonctionnement général

Pour aborder la problématique de la détection automatique de la subjectivité, nous exploitons deux méthodes qui ont fait leurs preuves ces dernières années. Le système que nous avons développé est décrit en figure 1.

La première consiste à utiliser la textométrie afin didentifier des descripteurs qui serviront de critères pour classifier les tweets (voir supra, 2.1). La seconde repose sur un apprentissage non-supervisé d'espaces vectoriels, à partir d'un autre corpus, dans lequel projeter les tweets avant de les classifier. Nous réalisons également un troisième système dont la sortie est déterminée par les deux premiers. Dans les trois cas, quelles que soient les premières étapes de traitement, la classification finale est supervisée, par appel à un logiciel d'apprentissage automatique.

2.1 Linguistique de corpus et textométrie

2.1.1 Le corpus DEFT et son annotation

Le corpus DEFT et son annotation sont caractérisés par quelques particularités qui conditionnent les résultats éventuels d'un système de classification automatique de tweets. Tout d'abord, il nous a semblé qu'une proportion non-négligeable du corpus est constituée de tweets générés automatiquement (boutons de partage situés sur les sites d'actualité). Ces tweets sont reconnaissables : à la mention d'un *via x* où *x* est le nom d'un site d'actualité ; à des énoncés tronqués ; à la présence d'URLs dans le corps du texte (par exemple : « *Forte hausse des raccordements éolien et photovoltaïque : Les raccordements d'éoliennes et de panneaux solaires... [URL]* »). Ils reprennent ainsi, dans un grand nombre de cas, les titres ou des chapeaux d'articles de presse et peuvent difficilement être classés en terme d'opinion ou de sentiments. En voici trois exemples tirés du corpus d'apprentissage :

— « *#Euthanasie : Un chien survit à une tentative d'euthanasie - 7sur7 [URL]* » : polarité positive, pas de classe générique (NULL) ;
— « *Un #chien survit miraculeusement à une tentative d'#euthanasie ! [URL]* » : polarité neutre, classe générique INFORMATION ;
— « *Un chien survit à une tentative d'euthanasie [URL] via @7sur7* » : polarité positive, classe générique OPINION, sous-classe VALORISATION.

De surcroît, comme nous le voyons dans l'exemple ci-dessus, les tweets peuvent être très similaires, et les divergences d'annotation importantes. L'annotation de certains tweets semble avoir été réalisée à partir de la simple présence d'un terme porteur d'une émotion ou d'une opinion spécifique (par exemple, le tweet « *Elles font **fureur**... Leur toucher doux, leur couvercle cristal, leur respect d'environnement ! URL* », classé dans la sous-classe COLERE). Enfin, il semble aussi que certains annotateurs font le choix de classer tous les tweets comportant un terme connoté positivement dans le domaine de l'écologie (« *renouvelable* », « *durable* », « *solaire* », etc.) dans la sous-classe VALORISATION, sans que ce choix soit unanime, comme nous le voyons dans ces tweets :

— « *BFM Business : Transition énergétique : le Syndicat des énergies renouvelables confiant [URL]* » polarité positive, pas de classe générique (NULL) ;
— « *Transition énergétique : le Syndicat des énergies renouvelables confiant [URL]* » : polarité positive, classe générique OPINION, sous-classe VALORISATION.

Sous-classe	**Corpus d'origine**	**Proportion (%)**	**Corpus réannoté**	**Proportion (%)**
ACCORD	**14**	3,11	**5**	1,11
AMOUR	0	-	0	-
APAISEMENT	1	0,22	3	0,67
COLERE	19	4,22	15	3,33
DEPLAISIR	3	0,67	6	1,33
DERANGEMENT	1	0,22	0	-
DESACCORD	5	1,11	8	1,78
DEVALORISATION	19	4,22	19	4,22
ENNUI	0	-	0	-
INSATISFACTION	1	0,22	0	-
MEPRIS	**8**	1,78	**19**	4,22
PEUR	**17**	3,78	**10**	2,22
PLAISIR	**0**	-	**5**	1,11
SATISFACTION	8	1,78	3	0,67
SURPRISE_NEGATIVE	1	0,22	2	0,44
SURPRISE_POSITIVE	1	0,22	0	-
TRISTESSE	1	0,22	2	0,44
VALORISATION	**78**	17,33	**30**	6,67
INFORMATION	**220**	48,89	**301**	66,89
NULL [pas de classe]	53	11,78	22	4,89
Total	450	-	450	-

TABLE 1 – Répartition du nombre de tweets par sous-classe au sein de notre échantillon réannoté

Ainsi, dans le cadre de cette campagne d'évaluation, nous avons voulu comparer l'annotation d'origine fournie par l'organisation de DEFT avec une annotation du même corpus réalisée par nos soins, en lançant une mini campagne d'étiquetage des tweets. A l'issue de la réannotation (en suivant le guide d'annotation de DEFT [2]) d'un échantillon de 450 tweets extraits au hasard du corpus d'entraînement, le taux de recoupement est d'environ 70%. Les disparités, surtout présentes pour les classes majoritaires INFORMATION et VALORISATION, s'expliquent par l'application de règles plus strictes dans notre annotation. Par exemple, nous avons choisi de ne classer dans VALORISATION que les tweets qui comportent un commentaire valorisant ou qui portent une marque d'engagement du rédacteur :

— « *Amateurs vegan y'a un super livre de @100vegetal qui va paraitre sur les fromages (j'en ai gouté 1, c'est trop bon) [URL]* » (dans le corpus d'origine : polarité neutre, classe générique INFORMATION) ;
— « *Je salue le courage des écologistes japonais qui luttent contra la chasse aux dauphins dans leur propre pays.* » (dans le corpus d'origine : polarité positive, classe générique SENTIMENT, sous-classe SATISFACTION).

Nous avons également classifié davantage de tweets sarcastiques dans la sous-classe MEPRIS : « *#Écologie #findumonde Cet homme, que dis-je ce héros, va nous sauver. #ohwait [URL]* ».

Ce processus de réannotation et de comparaison nous a permis d'évaluer l'homogénéité de certaines classes de manière qualitative. Le tableau 1 montre le nombre de tweets par sous-classe au sein de notre échantillon. En gras apparaissent les classes ou sous-classes qui présentent les plus fortes divergences : INFORMATION, VALORISATION et MEPRIS, mais aussi ACCORD, PEUR, PLAISIR et SATISFACTION. Ces deux dernières sous-classes se confondent facilement (avec APAISEMENT également). Par exemple, nous avons classifié le tweet *@Actuenviro : Transition énergétique : le maintien de Ségolène Royal rassure écologistes et industriels [URL]* dans la sous-classe APAISEMENT, alors qu'il apparaissait dans la sous-classe SATISFACTION : les deux annotations semblent correctes. Quant aux divergences pour la sous-classe PEUR, elles résultent de la frontière très ténue entre inquiétude et information (*Éolien - La possible extension des zones d'exclusion militaires provoque la crainte des professionnels : [URL]*).

2. https://deft.limsi.fr/2015/guideAnnotation.fr.php?lang=fr

2.1.2 Élaboration textométrique de critères de classification

L'élaboration textométrique des critères consiste à trouver des critères de classification linguistiquement explicables et suffisamment robustes pour servir de descripteurs aux méthodes d'apprentissage supervisé. L'analyse du corpus et le repérage des critères linguistiques ont été effectués avec deux logiciels textométriques : Lexico 3 (Salem *et al.*, 2003) et TXM (Heiden *et al.*, 2010) qui implémentent notamment les algorithmes de spécificités (Lafon, 1980) et de collocations (« Segments répétés » de Lexico 3) ainsi que les concordances qui nous ont permis le retour au texte et donc la vérification de la pertinence linguistique des critères.

An amont de l'analyse du corpus, nous avons effectué quelques prétraitements :
- les URLs ont été remplacés par la chaîne de caractère *URL*,
- les émoticônes ont été supprimés,
- les hashtags ont été considérés comme des mots simples (séparés du marqueur #),
- enfin, le corpus a été lemmatisé avec TreeTagger (Schmid, 1994). La lemmatisation, bien qu'elle fasse l'objet de débat en textométrie (Brunet, 2000) comme en analyse d'opinion (Pang *et al.*, 2002) nous a semblé un choix judicieux à cause de la particularité du corpus (textes courts avec peu de redondance de mots dans un texte) et de sa taille (en effet, nous avons constaté auparavant que les lemmes étaient plus performants sur de grands corpus).

Pour l'expérience nous avons utilisé trois types de critères :
- critères unitaires : choix des lemmes pertinents
- critères composites adjacents : choix des n-grammes de longueur variable de 2 à 6 unités
- cooccurrences textuels (dans la fenêtre d'un tweet) de 2 lemmes

Tous les critères sont sélectionnés selon trois principes : (i) leur caractère spécifique à une catégorie (ii) leur fréquence et (iii) leur pertinence linguistique.

Nous avons choisi les deux premiers types de critères selon le procédé suivant :

1. calcul des spécificités des lemmes isolés et de leur n-grammes (fonction « Segments Répétés » de Lexico 3) pour chaque catégorie ;
2. analyse des contextes d'apparition des lemmes spécifiques (au moyen de concordances textuelles) afin de s'assurer de leur pertinence textuelle et de l'unicité de leur fonction (les critères ayant une seule fonction et signification ont été privilégiés) ;

La sélection des cooccurrences a été réalisée comme suit :

1. calcul des paires de lemmes cooccurrents pour chaque tweet
2. calcul de spécificités de chaque cooccurrence pour toutes les catégories (avec le logiciel TXM)
3. sélection des cooccurrents sémantiquement interprétables (élimination des cooccurrents avec des mots-outils fréquents, choix des cooccurrents qui soit précisent un item déjà présent dans parmi les lemmes isolés soit apporte un nouveau critère sémantique).

2.1.3 Descripteurs linguistiques extraits

Nous présentons ici succinctement les principales catégories de critères qui ont servi à la classification des tweets. Nous exposons les critères obtenus avec la première méthode (méthode textométrique).

Nous distinguons quatre catégories de critères linguistiques : thymiques, dialogiques, dialectiques et thématiques.
- **Les critères thymiques** sont réputés intrinsèquement axiologiques et relèvent d'une vision classique de l'expression de la subjectivité. Pour catégoriser les tweets positifs, on trouve des marqueurs comme *bon, beau, intéressant, mieux, bien, positif, super, bravo, aimer*. Dans les tweets de polarité négative on recense les mots comme *mauvais, suspect, polémique, inquiéter, pire, mal, colère, foutre, con, merde, gueule*. Néanmoins, la proportion des marqueurs axiologiques reste relativement faible par rapport aux autres catégories, ce qui nous amène à penser que l'expression de la subjectivité est un phénomène complexe que l'on ne peut réduire à l'identification des marqueurs thymiques.
- **Les critères dialogiques** concernent la représentation des acteurs, le positionnement énonciatif et la distribution des rôles actanciels. Ils actualisent essentiellement les pronoms personnels, les pronoms possessifs et certaines entités nommées. On le trouve essentiellement dans les tweets de polarité négative ce qui peut s'interpréter comme un ancrage plus prononcé dans la présence du locuteur et dans l'interaction. Il s'agit essentiellement des pronoms comme *elle, lui, tu, te, on, me* et quelques entités nommées du domaine : *Ecologistes, Ségolène Royal, communiste*.

— **Les critères dialectiques** sont dédiés à la représentation du temps et du déroulement aspectuel, des structures argumentatives et de certaines modalités. Le vocabulaire la caractérisant est plus varié. Il peut s'agir de marqueurs de structuration, des verbes modaux, et des indicateurs rhétoriques (emphases, points d'interrogation, mots interrogatifs, etc.). Dans notre corpus, les critères dialectiques se trouvent principalement dans les tweets négatifs (*comment, pourquoi, ?, pourtant, ah*). Les tweets neutres se caractérisent par des ponctuations de phrase qui structurent le texte ; par conséquent on peut également les considérer comme dialectiques : *%, (,), ;, ...*
— **Les critères thématiques** sont les plus nombreux dans ce corpus. Ils caractérisent les différents thèmes abordés qui sont dans notre cas porteur d'une polarité. Les critères positifs sont liés à la sauvegarde de la nature et au développement des solutions alternatives pour l'énergie. Voici quelques exemples : *investir, réduire, soutenir, crowdfunding, géothermie, construire, développer, protection, cellule solaire, photovoltaïque, financement participatif, énergie positif, réduire CO2, réduire aéroport, développer renouvelable, créer écosystème*. Les critères négatifs expriment les problèmes écologiques : *en danger, disparition, crise, réchauffement climatique, espèce menacer, impasse climatique, écologie punitif, oiseau, neige, assassiner, mort, tuer, indifférence*. Les critères des tweets neutres (informationnelles) comportent quasi exclusivement les critères thématiques qui relatent les actualités. Les exemples sont les suivants : *publication, emploi, job, programme, panorama, consultation, rencontre, étudier, conférence, test, observatoire*.

2.1.4 Apprentissage automatique

Pour classer les tweets nous avons utilisé les algorithmes d'apprentissage supervisé. Nous en avons testé plusieurs, nous ne présentons ici que les résultats obtenus avec l'algorithmes de Machines à Vecteurs de Support (SMO) (Platt, 1998) intégré dans Weka (Hall *et al.*, 2009) qui a donné les meilleurs résultats. En amont des résultats sur le corpus de test, nous présentons les résultats obtenus sur le corpus d'apprentissage avec la validation croisée à dix plis en table 2.

Tâche	Macro-précision	Micro-précision
1	71,73	69,5
2.1	70,35	70,10
2.2	52,00	63,70

TABLE 2 – Résultats obtenus sur le corpus dapprentissage par validation croisée

2.2 Utilisation de l'apprentissage non supervisé

2.2.1 Calcul de l'espace vectoriel et projection des tweets

Dans le contexte de cette campagne d'évaluation, nous nous sommes aperçu que la taille des corpus d'entraînement est limitée et peu de ressources sont facilement disponibles pour le français. Afin de tester des approches qui limitent la dépendance au corpus en terme de vocabulaire, nous nous sommes tournés vers des algorithmes non supervisés. Les travaux récents de Mikolov *et al.* (2013) ont montré l'efficacité que l'on peut obtenir lors de l'utilisation de représentations de mots (ou d'expressions) dans des espaces vectoriels qui sont calculés selon leurs contextes. C'est l'objectif atteint par l'outil word2vec [3] qui a fait ses preuves dans d'autres domaines et que nous avons entraîné sur les corpus suivants :

Corpus	Mots (K)	Description
AFP	500 558	Dépêches AFP sur les années 2007-2013
Deft (train)	116	Corpus d'entraînement de DEFT
CoMeRe	568	Tweets de personnalités politiques (Longhi *et al.*, 2014)
Feelings	1 686	Extraction de tweets avec l'outil twitter-feelings
Hashtags	593	Extraction de tweets avec twython à partir de hashtags

TABLE 3 – Volumétrie et description du corpus d'entraînement de word2vec

3. https://code.google.com/p/word2vec/

Notre corpus dépasse les 500 millions de mots, dont la très large majorité est constituée de dépêches AFP. Une partie à été collectée à l'aide de l'outils twitter-feelings [4] ou par recherche de hashtags liés à l'écologie [5]. Le corpus a ensuite été lemmatisé avec TreeTagger (Schmid, 1994). De plus, les hashtag et les mentions sont introduits sous trois formes : tels quels ; sans leurs préfixes (@ ou #) ; segmentés selon la présence de majuscules. Le corpus est ensuite traité par word2vec afin d'apprendre des vecteurs de 500 composantes, sur une fenêtre contextuelle de 10 mots, en 20 itération (autres paramètres laissés par défaut).

Les tweets provenant du corpus d'entraînement ou du corpus de test sont prétraités avec les mêmes procédures, puis projetés dans l'espace vectoriel des lemmes créé par word2vec. Comme à chaque mot un vecteur est associé dans cet espace, la projection est une somme normalisée des mots présents dans chaque tweet (notre hypothèse étant que la longueur d'un message n'impacte pas les opinions / sentiments / émotions qui y sont présents). Nous ajoutons par ailleurs pour cette méthode la distance cosinus des mots du tweet avec chaque descripteur déterminé dans la partie 2.1.3.

2.2.2 Apprentissage automatique

Notre premier objectif est d'évaluer les performances obtenues par diverses approches. Pour ce faire, nous avons utilisé les algorithmes fournis par Weka (Hall *et al.*, 2009), ainsi que le filtre utilisé par défaut (StringToWordVector) permettant de convertir des textes sous formes de vecteurs de mots. Les tweets pouvaient être fournis : dans leur forme brute ; après une lemmatisation ; après projection dans l'espace vectoriel ; après projection dans l'espace vectoriel avec ajout des distances aux descripteurs. Nous avons ensuite évalué nos résultats sur le corpus d'entraînement grâce à l'outil proposé dans le cadre de la campagne, en réalisant nous-même une validation croisée à 10 plis.

Prétraitements	Algorithme	Macro-précision	Micro-précision
Tokens	ZeroR	15,04	45,12
	NaiveBayes	46,05	47,20
	NaiveBayesMultinomial	37,14	35,29
	SMO	58,34	59,46
Lemmes	NaiveBayes	47,41	48,19
	NaiveBayesMultinomial	38,39	35,71
	SMO	58,77	60,37
Vecs	NaiveBayes	49,71	51,15
	SMO	66,41	67,32
Vecs+Desc	NaiveBayes	50,09	51,42
	NaiveBayesMultinomial	62,67	60,89
	SMO	65,72	66,70

TABLE 4 – Comparaison des prétraitements et algorithmes pour T1

Les résultats obtenus par les algorithmes présentés en table 4 montrent la supériorité de l'algorithme SMO (Platt, 1998). Nous voyons également que, si la lemmatisation apporte assez peu, le passage aux représentations vectorielles apporte des gains très significatifs en performance. Effectivement, les représentations vectorielles permettent de limiter la dépendance aux corpus d'entraînement et donc de couvrir un vocabulaire qui n'est pas compris dans ce dernier. Par contre, l'ajout des distances aux descripteurs n'apportent pas plus (et font même légèrement baisser les performances).

2.3 Combinaison des méthodologies

Les méthodes sont combinées en regardant quelles catégories ont été bien annotées par chaque système sur une sous-partie du corpus de test que nous avons réannoté, comme cela a été décrit dans la partie 2.1.1. Ainsi, selon les résultats de chaque système, nous avons déterminé des heuristiques déterministes simples qui, à partir de la sortie des deux systèmes, prend une décision. Pour la tâche 1, les règles sont les suivantes :

4. https ://github.com/cblavier/twitter-feelings

5. Liste des hashtags : #ecologie #Ecologie #écologie #Environnement #Biodiversité #DD #biodiversité #énergie #solaire #Energie #environnement #énergies #EELV #Animaux #Durable #climat

— choisir neutre si telle est la sortie du système word2vec,
— choisir positif ou négatif si telle est la sortie de la méthode textométrique,
— sinon, mettre neutre.

Pour la tâche 2.1, nous donnons la priorité par classe, quel que soit le système considéré, dans l'ordre suivant : sentiment, information, émotion, opinion. Nous adoptons le même principe pour la tâche 2.2 avec l'ordre suivant : DÉPLAISIR, TRISTESSE, INSATISFACTION, PLAISIR, DÉSACCORD, DÉRANGEMENT, AMOUR, SATISFACTION, MÉPRIS. Notons que cette combinaison n'a été réalisée qu'à titre expérimental, ce qui explique son peu de sophistication (utiliser un apprentissage automatique à ce niveau aurait probablement été plus efficace).

3 Résultats

Les résultats obtenus sur le corpus de test sont présentés en table 5. Pour la tâche 1, nous constatons que la méthode textométrique obtient de meilleurs résultats que word2vec, et que la combinaison de ces deux systèmes nous permet d'obtenir des résultats très proches du meilleur système de la campagne. Pour les tâches 2.1 et 2.2, word2vec obtient des résultats meilleurs que la méthode textométrique, approchant une nouvelle fois le meilleur système pour la tâche 2.2, la combinaison n'obtenant de bons résultats que pour la tâche 2.1.

Tâche	Textométrie	word2vec	Combinaison	Max. DEFT
1	71,09	69,17	73,44	73,60
2.1	56,22	57,19	57,53	61,29
2.2	29,23	33,72	30,42	34,68

TABLE 5 – Résultats DEFT 2015

Ces résultats sont satisfaisants au regard des meilleurs systèmes de la campagne et montrent bien les avantages et inconvénients de chaque méthode utilisée. La méthode textométrie permet effectivement, pour une classification binaire, de s'appuyer sur des descripteurs plutôt que sur les mots du corpus afin de construire des représentations pertinentes des messages pour l'apprentissage automatique. Pour des tâches demandant une classification plus fine, la méthode textométrique a montré ici des limites. Les représentations vectorielles donnent de meilleurs résultats. Nos expériences et évaluations sur le corpus d'entraînement nous font suspecter une forte spécificité des descripteurs pour ces tâches, et donc un sur-apprentissage, ce que pallie word2vec en évitant de s'appuyer sur les mots eux-mêmes, mais sur leur projection dans un espace continu.

Conclusion

La campagne d'évaluation DEFT 2015 sur l'annotation subjective de tweets nous permet de mener des travaux dans deux directions. La première porte sur l'annotation elle-même et vise à déterminer comment il est possible d'extraire des indices permettant (pour un humain ou un algorithme) de déterminer quelle classe attribuer à un texte court. Si nos conclusions à cet égard sont encore parcellaires, nous nous apercevons de la difficulté de la tâche et de ses variabilités. La seconde direction vise à fonder expérimentalement les méthodes adéquates pour construire des systèmes qui classent automatiquement ces tweets. Nous expérimentons une méthode textométrique et mettons ici en avant les bonnes performances qu'elle obtient en combinaison avec un apprentissage automatique. Par ailleurs, nous la confrontons également aux représentations vectorielles, qui montrent également leur intérêt, en particulier lorsque les catégories sont nombreuses et le corpus d'entraînement de taille limité. Comme les résultats obtenus le montrent, combiner le deux permet d'obtenir des résultats très compétitifs, une perspective que nous envisageons d'approfondir dans nos travaux futurs.

Références

BRUNET E. (2000). Qui lemmatise dilemme attise. *Lexicometrica*, **2**.

EENSOO E. & VALETTE M. (2012). Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 2, p. 357–374, Grenoble.

EENSOO E. & VALETTE M. (2014a). Approche textuelle pour le traitement automatique du discours évaluatif. *A. Jackiewicz, (éd.), Études sur l'évaluation axiologique, Langue française*, (184), 107–122.

EENSOO E. & VALETTE M. (2014b). Sémantique textuelle et tal : un exemple d'application à l'analyse des sentiments. *D. Ablali, S. Badir, D. Ducard, Eds., Documents, textes, œuvres, Presses Universitaires de Rouen, Collection Rivages linguistiques*.

EENSOO E. & VALETTE M. (2015). Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité. In *Actes de la conférence TALN 2015*.

GHORBEL H. & JACOT D. (2011). Further experiments in sentiment analysis of french movie reviews. In E. MUGELLINI, P. SZCZEPANIAK, M. PETTENATI & M. SOKHN, Eds., *Advances in Intelligent Web Mastering 3*, volume 86 of *Advances in Intelligent and Soft Computing*, p. 19–28. Springer Berlin / Heidelberg. 10.1007/978-3-642-18029-3_3.

HALL M., EIBE F., HOLMES G., PFAHRINGER B., REUTEMANN P. & WITTEN I. H. (2009). The weka data mining software : An update. *SIGKDD Explorations*, **11**(1).

HEIDEN S., MAGUÉ J. P. & PINCEMIN B. (2010). Txm : Une plateforme logicielle open-source pour la textométrie conception et développement. In S. BOLASCO, Ed., *Actes de la conférence JADT 2010*, volume 2, p. 1021–1032.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, **1**, 127–165.

LAFON P. (2981). Analyse lexicométrique et recherche des cooccurrences. *Mots*, (3), 95–148.

LONGHI J., MARINICA C., BORZIC B. & ALKHOULI A. (2014). Polititweets, corpus de tweets provenant de comptes politiques influents. *corpus]. In Chanier T.(ed) Banque de corpus CoMeRe. Ortolang. fr : Nancy. http ://hdl. handle. net/11403/comere/cmr-polititweets*.

MAUREL S. & DINI L. (2009). Exploration de corpus pour l'analyse des sentiments. In *Actes de DEFT'09 ń DÉfi Fouille de Textes ż, Atelier de clôture*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

PANG P., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? sentiment classification using machine learning techniques. In *In Proceedings of EMNLP*, p. 79–86.

PLATT J. (1998). Machines using sequential minimal optimization. In B. SCHOELKOPF, C. BURGES & A. SMOLA, Eds., *Advances in Kernel Methods - Support Vector Learning*.

RASTIER F. (2001). *Arts et sciences du texte*. Presses Universitaires de France.

SALEM A., LAMALLE C., MARTINEZ W., FLEURY S., FRACCHIOLLA B., KUNCOVA A. & MAISONDIEU A. (2003). *Lexico3 Outils de statistique textuelle. Manuel d'utilisation*. Syled-CLA2T, Université Sorbonne Nouvelle.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

TURNEY P. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics (ACL)*, p. 417–424.

VERNIER M., MONCEAUX L. & DAILLE B. (2009a). Deft'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique. In *Actes de l'atelier de clôture de la 5ème édition du Défi Fouille de Textes*.

VERNIER M., MONCEAUX L., DAILLE B. & DUBREIL E. (2009b). Catégorisation des évaluations dans un corpus de blogs multi-domaine. *Revue des nouvelles technologies de l'information (RNTI)*, p. 45–70.

YI J., NASUKAWA T., BUNESCU R. & NIBLACK W. (2003). Sentiment analyzer : Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, p. 427–, Washington, DC, USA : IEEE Computer Society.

YU H. & HATZIVASSILOGLOU V. (2003). Towards answering opinion questions : separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, p. 129–136, Stroudsburg, PA, USA : Association for Computational Linguistics.