Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (DEFT)

Thierry Hamon, Amel Fraisse, Patrick Paroubek, Pierre Zweigenbaum, Cyril Grouin

Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (DEFT)

Thierry Hamon[1,2] Amel Fraisse[1] Patrick Paroubek[1] Pierre Zweigenbaum[1] Cyril Grouin[1]
(1) LIMSI-CNRS, Campus universitaire d'Orsay, Rue John von Neumann, Bât 508, 91405 Orsay
(2) Université Paris 13, Villetaneuse, France
prenom.nom@limsi.fr

Résumé. L'édition 2015 du défi fouille de texte (DEFT) porte sur la fouille d'opinion et l'analyse des sentiments et des émotions dans les messages postés sur Twitter en relation avec la thématique du changement climatique. Trois tâches ont été proposées : (i) déterminer la polarité globale des tweets, (ii) identifier les classes génériques (opinion, sentiment, émotion, information) et spécifiques (parmi 18 classes) de ces tweets, et (iii) analyser la source, la cible et l'expression porteuse d'opinion, de sentiment ou d'émotion. Douze équipes ont participé. Les meilleurs résultats, en macro-précision, sont de 0,736 (polarité), 0,613 (classes génériques) et 0,347 (classes spécifiques). Aucun participant n'a soumis de données pour la dernière tâche. Les méthodes utilisées reposent majoritairement sur des approches par apprentissage statistique supervisé (SVM, Naïve Bayes, réseaux neuronaux, PPMC), et utilisent de nombreux lexiques d'opinions (ANEW, Casoar, Emotaix, Feel, Lidilem) et de polarités (Polarimots) comme traits.

Abstract.

Analysis of emotion, sentiment and opinion within tweets. Presentation and results of the 2015 DEFT text mining challenge

The 2015 DEFT text mining challenge focused on opinion mining, emotion and sentiment analysis of messages from Twitter, on the climate change thematic. Three tasks were proposed : (i) determine the general polarity of tweets, (ii) identify generic classes (opinion, sentiment, emotion, information) and specific classes (among 18 classes) from these tweets, and (iii) analyze source, target, and opinion, sentiment, emotion focus. Twelve teams participated. The best results, in terms of macro-precision, are of 0.736 (polarity), 0.613 (generic classes) and 0.347 (specific classes). No run was submitted for the last task. The methods used by the participants mainly rely on statistical machine learning approaches (SVM, Naïve Bayes, neural network, PPMC), using several opinion lexicon (ANEW, Casoar, Emotaix, Feel, Lidilem) and polarity lexicon (Polarimots) as features.

Mots-clés : Fouille d'opinion, analyse d'émotions, analyse de sentiments, réseaux sociaux, campagne d'évaluation.

Keywords: Opinion mining, Emotion analysis, Sentiment Analysis, Social network, NLP Challenge.

1 Introduction

Le défi DEFT est un atelier annuel d'évaluation francophone en fouille de textes. Les thématiques abordées relèvent du domaine expérimental et visent à vérifier la faisabilité des tâches proposées au moyen des méthodes disponibles.

La fouille d'opinion constitue une activité qui a déjà été proposée lors de deux précédentes éditions de DEFT. En 2007, nous proposions de travailler sur la retranscription de débats parlementaires pour déterminer l'opinion véhiculée par le message (Grouin *et al.*, 2007, 2009b). Il s'agissait donc pour les participants de travailler sur des textes correctement écrits, composés de phrases assez longues. L'édition 2009 a permis de comparer les opinions exprimées dans deux corpus différents, un corpus de débats parlementaires d'une part, et un corpus d'articles de journaux (éditoriaux, articles d'analyse et de débat, articles de fond) d'autre part (Grouin *et al.*, 2009a). En dehors de l'identification du caractère objectif ou subjectif d'un texte, nous avons proposé aux participants de cette édition d'identifier les expressions porteuses d'opinion. Pour cette tâche, faute de disposer de moyens pour constituer la référence, nous avions évalué les résultats des participants sur la base des annotations communes aux différentes soumissions.

Pour cette onzième édition, nous proposons de travailler sur l'analyse de l'opinion, des sentiments et des émotions dans des tweets rédigés en français. Contrairement aux précédentes éditions, nous proposons cette année un ensemble de tâches permettant de traiter la fouille d'opinion de manière complète, tant du point de vue de la polarité et de l'opinion globale d'un message que du point de vue de l'identification de la source, de la cible et de l'expression porteuse d'opinion à l'intérieur de chaque message. Le corpus proposé a fait l'objet d'une annotation complète par des annotateurs humains, en fonction de principes définis dans un guide d'annotation (Fraisse & Paroubek, 2014).

2 Corpus

2.1 Présentation

Le corpus se compose de 15 000 messages postés sur le réseau social Twitter, en relation avec la thématique du changement climatique. Ces messages, appelés « *tweets* » et d'une longueur maximale de 140 caractères, sont caractérisés par : (i) des phrases courtes, (ii) l'utilisation d'abréviations génériques ou propres à l'Internet (*MDR, STP, klk1, kan, dla*), (iii) un style littéraire plus ou moins familier selon l'émetteur du message (« *ils vous bananes au calme* »), (iv) la présence d'émoticônes (ou *smiley* : `xD`), et (v) la présence de mots-clés spécifiques au réseau Twitter : des *hashtags* ou *mot-dièses* commençant par le symbole dièse pour marquer une thématique (`#Irak`, `#NoControlDay`) ou un état d'esprit présenté de manière sarcastique (`#FeedTheTroll`), et des noms d'utilisateurs commençant par le symbole arobase (`@CNRS`).

2.2 Annotation

2.2.1 Procédure

Les tweets ont été annotés par deux annotateurs. Une première étape de double annotation portant sur 500 messages a été réalisées. Elle a permis aux annotateurs de se familiariser avec le guide d'annotation. La suite des messages a été annotée en simple annotation. Nous avons calculé les accords inter-annotateurs sur les 500 messages annotés en double.

Le tableau 1 renseigne des probabilités observées (Pa), probabilités attendues (Pe) et des valeurs de Kappa et de coefficient de Dice calculées en première approximation pour l'accord sur le cardinal (nombre d'instances trouvées par un annotateur independamment des positions respectives de ces instances dans les documents) pour chacune des catégories possibles d'annotation, sur le corpus de 500 tweets annotés en double. Le choix de calculer l'accord sur le cardinal des catégories d'annotation au lieu de le faire sur les annotations est le résultat de contraintes temporelles qui n'ont pas permis de développer des mesures de Kappa adaptées aux annotations fines de la tâche (iii) pour prendre en compte de manière raisonnable les différences de frontières d'empan de texte qui aboutissent à des valeurs artificiellement faibles de kappa si l'on considère une égalité stricte au niveau du caractère.

Si globalement, les accords inter-annotateurs témoignent d'une absence d'accord entre annotateurs avec une valeur de $\kappa < 0,8$ (Artstein & Poesio, 2008), précisons que cette valeur tient compte de l'ensemble des catégories et relations utilisées dans le schéma d'annotation. Nous observons qu'il n'est pas possible de dégager un consensus entre les deux annotateurs sur certaines catégories, en particulier la polarité *Positif* ($\kappa = 0,18$), l'identification de la cible ($\kappa = 0,39$) et plusieurs classes spécifiques (du plus consensuel au moins consensuel, pour des valeurs de $\kappa < 0,8$: *Accord, Valorisation, Mépris, Tristesse, Apaisement, Satisfaction, Dévalorisation, Colère*).

2.2.2 Guide d'annotation

Chaque message est annoté au moyen de sept groupes, dont un comprend 19 catégories fines, et de cinq relations [1].

Groupes Les sept groupes utilisés sont les suivants :

– SOURCE : groupe de mots qui référence l'auteur de l'expression d'opinion, sentiment, émotion (OSEE : *Opinion Sentiment Emotion Expression*). On annote la mention explicite de la source. Celle-ci doit être la plus large possible incluant ses modifieurs, ses circonstants, ses multiples apposés, ses relatives, mais également les conjonctions

1. Le guide utilisé est accessible à l'adresse suivante : `https://deft.limsi.fr/2015/guideAnnotation.fr.php?lang=fr` Nous invitons le lecteur à consulter le guide d'annotation pour accéder aux exemples associés à chacun des groupes, catégories et relations.

	OSEE_GLOBALE	RELATIONS					SOURCE, CIBLE, EXPRESSION					POLARITE		
		DIT	SUR	MOD	NEG	RECEPTEUR	Source	Cible	Modifieur	Négation	Destinataire	Positif	Négatif	Inconnu
Pa(==)	0,76	1,00	1,00	1,00	1,00	1,00	0,95	0,83	0,99	0,99	1,00	0,94	0,93	1,00
Pa	0,76	1,00	1,00	1,00	1,00	1,00	0,95	0,83	0,99	0,99	1,00	0,94	0,93	1,00
Pe	0,24	0,75	0,35	0,94	0,94	0,98	0,75	0,39	0,94	0,94	0,98	0,92	0,82	1,00
Kappa	0,69	1,00	1,00	1,00	1,00	1,00	0,79	0,72	0,82	0,80	1,00	0,18	0,62	1,00
Dice	0,76	1,00	1,00	1,00	1,00	1,00	0,95	0,83	0,99	0,99	1,00	0,94	0,93	1,00

	CLASSES SPECIFIQUES																			
	Amour	Plaisir	Apaisement	Surprise_positive	Satisfaction	Accord	Valorisation	Désaccord	Dévalorisation	Insatisfaction	Mépris	Colère	Tristesse	Déplaisir	Ennui	Peur	Dérangement	Surprise_négative	Instruction_Demande
Pa(==)	1,00	0,99	1,00	1,00	0,99	0,97	0,89	0,99	0,98	1,00	0,98	0,99	1,00	1,00	1,00	0,99	1,00	1,00	0,97
Pa	1,00	0,99	1,00	1,00	0,99	0,97	0,89	0,99	0,98	1,00	0,98	0,99	1,00	1,00	1,00	0,99	1,00	1,00	0,97
Pe	1,00	0,97	0,99	1,00	0,98	0,90	0,66	0,92	0,97	1,00	0,94	0,98	0,99	0,99	1,00	0,91	1,00	1,00	0,74
Kappa	0,00	0,71	0,66	1,00	0,50	0,73	0,69	0,83	0,37	1,00	0,69	0,36	0,67	0,80	1,00	0,86	0,00	1,00	0,87
Dice	1,00	0,99	1,00	1,00	0,99	0,97	0,89	0,99	0,98	1,00	0,98	0,99	1,00	1,00	1,00	0,99	1,00	1,00	0,97

TABLE 1 – Taux d'accord inter-annotateur (Kappa, Dice) calculés sur les 500 tweets annotés en double pour l'accord sur les cardinaux de catégories d'annotation

de modifieurs de toutes sortes, y compris de relatives, de manière à avoir le maximum d'information sémantique (*En tant que cuisinier amateur qui a de l' expérience , je n' aime pas vraiment les pâtes .*) ;

- CIBLE : mention explicite de la cible la plus large possible incluant ses modifieurs, ses circonstants, ses multiples apposés, ses relatives, voire aussi les conjonctions de modifieurs de toutes sortes y compris de relatives afin d'avoir le maximum d'information sémantique. Lorsque l'OSEE porte sur plusieurs cibles, chaque cible est identifiée dans un empan de texte distinct (*Les lynx , les loups , les tortues sont des éspéces protégées .*) ;
- NÉGATION : marqueurs de négation (*ne pas, ne plus,* etc.) (*Le serpent n' est pas une espèce protégée .*) ;
- MODIFIEUR : tout modifieur (*le plus*, etc.) (*La Sardine l' un des poissons les plus en danger en Méditerranée http :*) ;
- DESTINATAIRE : mention explicite du destinataire. Ce groupe sera essentiellement utilisé dans le cas ou l'expression d'opinion, sentiment, émotion est adressée à une entité *(personne, organisation, etc.)* (*Mme Ségolène Royale vous êtes priée de respecter les loups .*) ;
- EXPRESSION D'OPINION, SENTIMENT, ÉMOTION (OSEE) : empan de texte dont la valeur sémantique correspond à l'expression d'opinion, de sentiment ou d'émotion. Cette expression est annotée au moyen de l'une des 19 catégorie sémantiques fines [2] :
  - Classes sémantiques affectives fines : ces classes relèvent de trois grandes catégories :
    - Opinions (intellectif) : cette catégorie contient deux classes sémantiques positives (*Accord* et *Valorisation*) de deux classes sémantiques négatives (*Désaccord* et *Dévalorisation*) ;
    - Sentiments (affectif-intellectif) : cette catégorie contient une classe sémantique positive (*Satisfaction*) et une classe sémantique négative (*Insatisfaction*) ;
    - Émotions (affectif) : cette catégorie contient 4 classes sémantiques positives (*Plaisir, Apaisement, Amour* et *Surprise positive*) et 8 classes sémantiques négatives (*Déplaisir, Dérangement, Mépris, Tristesse, Peur, Colère, Ennui* et *Surprise négative*).

    Une dernière classe sémantique correspond aux instructions et aux informations (*instruction*).
  - Classes génériques de polarité :

2. Pour des raisons de lisibilité, les définitions et exemples de chacune de ces 19 catégories sont donnés dans l'annexe A.

  - *Négatif* (expressions d'opinion/sentiment/émotion qui ont une polarité négative et dont il est difficile d'identifier avec certitude la classe sémantique exacte),
  - *Positif* (expressions d'opinion/sentiment/émotion qui ont une polarité positive et dont il est difficile d'identifier avec certitude la classe sémantique exacte).
- OSE_GLOBALE : catégorie sémantique globale et générale du message. Dans le cas où le message contient plus d'une catégorie sémantique, il faut indiquer la catégorie sémantique dominante du message. Dans le cas contraire, l'OSE_GLOBALE aura la même valeur que le groupe d'expression d'opinion, sentiment, émotion du message.

Relations Les cinq relations considérées sont les suivantes :
- DIT : permet de mettre en rapport la SOURCE avec l'OSEE. Elle met toujours en rapport 2 groupes (*Je n' aime pas vraiment les pâtes* .) ;
- SUR : permet de mettre en rapport l'expression d'opinion, sentiment, émotion avec la CIBLE. Elle met toujours en rapport 2 groupes (*Je n' aime pas vraiment les pâtes* .) ;
- MOD permet de mettre en rapport les éventuels modifieurs de l'OSEE. Cette relation va mettre en rapport une forme (e.g., modifieur) avec le groupe d'expression d'opinion, sentiment, émotion ;
- NEG : permet de mettre en rapport les éventuels marqueurs de négation avec l'OSEE dont ils modifient la sémantique. Chaque marqueur suscite la création d'une relation NEG (notamment les deux éléments de la négation *ne* et *pas*) ;
- RECEPTEUR : permet de mettre en rapport le groupe de l'OSEE avec le groupe DESTINATAIRE.

2.2.3 Annotations de référence

En fonction de ces annotations fines utilisées comme annotations de référence pour la tâche 3 (voir section 3.1), nous avons dégagé automatiquement les valeurs de référence des tâches plus génériques (tâches 1, 2.1 et 2.2) au moyen d'une simple correspondance entre catégories fines et valeurs génériques (voir tableau 2). Les annotations de référence des premières tâches n'ont donc pas été directement établies par les annotateurs humains, mais inférées des annotations humaines de la troisième tâche.

Catégorie fine (T3)	Type (T2)	Polarité (T1)
Accord, Valorisation	opinion	+
Désaccord, Dévalorisation		-
Satisfaction	sentiment	+
Insatisfaction		-
Plaisir, Apaisement, Amour, Surprise positive	émotion	+
Déplaisir, Dérangement, Mépris, Surprise négative, Peur, Colère, Ennui, Tristesse		-
Instruction, Information	information	=

TABLE 2 – Correspondance entre catégories fines, type et polarité

Le tableau 3 renseigne de la distribution des annotations pour chacune des catégories proposées dans chaque tâche, entre corpus d'apprentissage et de test. Notons que l'équilibre entre les deux corpus est respecté pour chaque catégorie, à l'exception de la catégorie *Sentiment* de la tâche 2.1, sous-représentée dans le corpus d'apprentissage, et qui induit un déséquilibre sur les autres catégories. Ce déséquilibre provient de la difficulté de maîtriser la distribution entre catégories, à la fois entre corpus et entre tâches, lorsque le même corpus est utilisé pour plusieurs tâches.

3 Présentation du défi

3.1 Tâches proposées

Nous avons proposé trois tâches complètes autour des émotions, sentiments et opinions exprimées dans les messages postés sur Twitter. Les différentes tâches proposent un niveau d'analyse du plus global *(polarité, classes génériques)* au plus fin *(identification des classes spécifiques ; reconnaissance des expression porteuse d'opinion, cible et source dans le texte du tweet)*. L'ensemble de ces tâches permet de couvrir une large part des travaux possibles en matière d'analyse des

Tâche	Catégorie	Apprentissage	Test
T1	+	2464 (31,08%)	1057 (31,28%)
	-	1894 (23,89%)	804 (23,79%)
	=	3571 (45,04%)	1518 (44,92%)
T2.1	Emotion	826 (12,23%)	351 (10,39%)
	Information	3571 (52,87%)	1518 (44,92%)
	Opinion	2275 (33,68%)	973 (28,80%)
	Sentiment	82 (1,21%)	537 (15,89%)
T2.2	Accord	154 (4,84%)	67 (4,92%)
	Amour	8 (0,25%)	4 (0,29%)
	Apaisement	9 (0,28%)	5 (0,37%)
	Colère	210 (6,60%)	87 (6,39%)
	Déplaisir	47 (1,48%)	21 (1,54%)
T2.2 (suite)	Dérangement	13 (0,41%)	6 (0,44%)
	Désaccord	216 (6,79%)	92 (6,76%)
	Dévalorisation	401 (12,60%)	170 (12,49%)
	Ennui	4 (0,13%)	2 (0,15%)
	Insatisfaction	9 (0,28%)	5 (0,37%)
	Mépris	176 (5,53%)	75 (5,51%)
	Peur	274 (8,61%)	114 (8,38%)
	Plaisir	35 (1,10%)	15 (1,10%)
	Satisfaction	73 (2,29%)	32 (2,35%)
	Surprise_négative	10 (0,31%)	4 (0,29%)
	Surprise_positive	4 (0,13%)	2 (0,15%)
	Tristesse	36 (1,13%)	16 (1,18%)
	Valorisation	1504 (47,25%)	644 (47,32%)

TABLE 3 – Distribution des annotations par catégories et par tâches pour chaque corpus

émotions, sentiments et opinions appliquée aux messages courts postés sur les réseaux sociaux. Chaque participant était libre de choisir les tâches auxquelles il souhaitait participer..

Pour illustrer chacune de ces tâches, nous prenons appui sur le tweet suivant pour illustrer les annotations réalisées.

Energie renouvelable pleine de promesses, la géothermie souffre pourtant d'un manque de visibilité...

FIGURE 1 – Extrait du corpus (identifiant : 519507340304084992)

3.1.1 Tâche 1 – Polarité des tweets

La première tâche vise à détecter la polarité des tweets parmi trois valeurs possible : *positif* (+), *neutre ou mixte* (=), et *négatif* (−). La catégorie *neutre ou mixte* renvoie aussi bien aux messages présentant une polarité neutre (ni positif, ni négatif), que ceux présentant les deux polarités en même temps (un sentiment positif et un sentiment négatif). Le tweet présenté en exemple est classé *négatif* pour la tâche 1.

3.1.2 Tâche 2 – Classe des tweets

Cette tâche vise une classification fine des tweets. Nous avons divisé cette tâche en deux sous-tâches.

Tâche 2.1 – Classe générique Cette première sous-tâche vise l'identification de la classe générique de l'information exprimée dans le tweet, parmi quatre classes : *opinion, sentiment, émotion, information*. Le tweet présenté en exemple est classé *émotion* pour la tâche 2.1.

Tâche 2.2 – Classe spécifique Cette deuxième sous-tâche vise l'identification de la classe spécifique de l'opinion, du sentiment, ou de l'émotion exprimée, parmi dix-huit classes : *accord, amour, apaisement, colère, déplaisir, dérangement, désaccord, dévalorisation, ennui, insatisfaction, mépris, peur, plaisir, satisfaction, surprise négative, surprise positive, tristesse, valorisation*. Le tweet présenté en exemple est classé *déplaisir* pour la tâche 2.2.

3.1.3 Tâche 3 – Source, cible et expression d'opinion

Cette dernière tâche vise à analyser plus précisément les opinions, du point de vue de l'expression porteuse de l'opinion, de la source (l'émetteur) et de la cible (le récepteur).

Sur le tweet d'exemple, les portions suivantes sont annotées comme suit :

– Entités : « *Energie renouvelable* » est marquée *cible*, « *pleine de promesses* » est marquée *valorisation*, « *souffre* » est marqué *déplaisir*, et « *manque de visibilité* » est marqué *négatif* ;
– Relations : des relations DIT entre « *pleine de promesses* » et « *Energie renouvelable* », et entre « *manque de visibilité* » et « *souffre* ».

3.2 Organisation

A l'image des campagnes d'évaluation précédentes, cette édition s'est déroulée en deux temps. La première phase permet aux participants de développer et d'entraîner leurs systèmes à partir des données annotées qui leur sont fournies. Les inscriptions (remplissage d'un simple formulaire sur internet) et l'accès aux données annotées d'entraînement ont été autorisés à partir du 16 février 2015. Nous relevons que les inscriptions se sont réparties entre le 16 février et le 1^{er} mai, avec une majorité d'inscriptions au mois de mars. Au total, 23 équipes se sont inscrites, dont deux issues d'industriels (Proxem et Synapse Développement) et une équipe académique de l'Université Technique de Moldavie (TU Moldova).

La deuxième phase permet aux participants d'appliquer les méthodes qu'ils ont développées pendant la phase d'entraînement sur le corpus de test. Cette phase de test s'est déroulée du 4 au 10 mai 2015. Chaque participant a bénéficié d'une fenêtre de trois jours (définie par chaque équipe selon ses préférences) entre l'accès aux données de test et la soumission des résultats produits par son système. Nous avons reçu les soumissions de 12 équipes, chaque équipe pouvant soumettre jusqu'à trois sorties différentes de leur système, pour chacune des tâches proposées.

Conformément aux règles d'accès à Twitter et d'utilisation des tweets, lors des phases de développement et de test, nous avons fourni aux participants les identifiants des tweets et les outils permettant de constituer le corpus par eux-mêmes.

Sur la tâche 1, nous avons reçu un total de 27 soumissions correspondant à l'ensemble des 12 équipes ayant participé au défi. Sur la tâche 2.1, nous avons reçu 24 soumissions pour 9 équipes, et 21 soumissions pour 7 équipes sur la tâche 2.2. Aucune équipe n'a participé à la tâche 3.

Les évaluations des soumissions effectuées ainsi que les annotations de référence sur le corpus de test ont été communiqués aux participants entre le 14 et 18 mai 2015. Pour chacune des tâches proposées, chaque participant a eu accès à ses résultats individuels (pour l'ensemble des soumissions effectuées) ainsi qu'à des éléments de comparaison calculés sur la meilleure soumission de chaque équipe (moyenne, médiane, écart-type, valeurs minimum et maximum), le classement final n'étant dévoilé que le jour de l'atelier de clôture du défi (voir section 5). Pendant cette période, les participants ont été invités à vérifier les résultats calculés et à se prononcer sur le cas d'équipes ayant soumis des résultats avec des noms de catégories erronées sur la tâche 2.2 (voir section 3.1, utilisation de la catégorie « *instruction* » au lieu de « *information* », ou de versions abrégées « i », « o », « e » et « s » au lieu de « *information* », « *opinion* », « *émotion* » et « *sentiment* ») dont la correction modifie les résultats et le classement final.

3.3 Évaluation

Les résultats des tâches 1, 2.1 et 2.2 ont été évalués en termes de macro-précision (formule 1) (Manning & Schütze, 2000).

$$\text{Macro-précision} = \frac{\sum_{i=1}^{n} \left(\frac{\text{vrais positifs}(i)}{\text{vrais positifs}(i)+\text{faux positifs}(i)} \right)}{n} \quad (1)$$

Tous les tweets devant se voir attribué une catégorie, nous avons choisi de pénaliser fortement les systèmes ne prenant pas de décision ou proposant une catégorie non attendue initialement.

Le fait que les participants devaient constituer les corpus par eux-mêmes a conduit à une difficulté supplémentaire lors de l'évaluation. En effet, les tweets pouvant être supprimés par leur auteurs à n'importe quel instant, il était possible qu'un participant puisse récupérer un tweet qui ne serait plus disponible plus tard, pour d'autres participants. Il fallait nous assurer que tous les participants soient évalués sur le même ensemble de tweets. A la fin de la phase de test, nous avons donc identifiés les tweets supprimés. Ainsi, deux tweets avaient été supprimés pendant la période de test. Nous avons également constaté que plusieurs participants ont eu des difficultés, probablement techniques, à accéder à deux autres tweets. Nous avons choisi de ne pas prendre en compte ces quatre tweets dans l'évaluation finale.

4 Méthodes des participants

Chaîne de traitements TextAnalyst La société Synapse (équipe 19) a traité les corpus au moyen de la chaîne de traitements TextAnalyst, composée de lexiques et de plusieurs modules qui s'enchaînent en cascade (analyse syntaxique avec Cordial, détection des expressions d'opinion, détection et application d'opérateurs (négation, intensification, modalités), et calcul de l'opinion globale du tweet fondé sur le modèle parabolique) (Chardon *et al.*, 2015).

Apprentissage statistique La majorité des systèmes utilisés par les participants repose cependant sur des approches par apprentissage statistique supervisé. Les principaux algorithmes utilisés sont : SVM (LINA/Dictanova, équipe 6 ; LIRMM, équipe 17 ; LINA–Dimeco, équipe 25), Naïve Bayes (IRISA, équipe 14 ; LIMSI, équipe 23), réseau de neurones (IRISA, équipe 14 ; Proxem, équipe 15), ou encore un algorithme de prédiction par correspondance partielle : PPMC (TU Moldova, équipe 22). Le LIF (équipe 3) a utilisé plusieurs modèles probabilistes dont les résultats ont été fusionnés au moyen d'une procédure de vote.

Parmi les traits utilisés, l'équipe LINA–Dimeco (Lejeune & Dumonceaux, 2015) a émis l'hypothèse que le style utilisé dans un tweet reflète l'émotion de l'émetteur, et a donc mobilisé des critères stylométriques sur les caractères et mots du message pour en déterminer les émotions, opinions et sentiments. Le LIRMM (Abdaoui *et al.*, 2015) a par ailleurs pris en compte les patrons syntaxiques. Plusieurs équipes n'ont pris en compte que les descripteurs les plus discriminants, notamment sur le plan sémantique. Le LIMSI (Morlane-Hondère & D'hondt, 2015) a appliqué une méthode de sélection d'attributs par évaluation du gain d'information (fonction `InfoGainAttributeEval` de Weka (Witten & Frank, 2005)) afin de restreindre l'analyse des tweets aux 450 n-grammes de mots les plus discriminants. De manière similaire, l'INaLCO (équipe 2) a dégagé des descripteurs sémantiques au moyen d'une analyse textométrique. Plusieurs équipes ont également étudié la polarité des tweets, soit par la présence de négations (IRIT/LIMSI, Synapse), soit par des listes de termes polarisés positifs et négatifs (LIMSI). Notons que l'équipe TU Moldova (Bobicev, 2015) a réalisé plusieurs expériences fondées sur le traitement des caractères uniquement ou des mots uniquement, avec et sans normalisation. Les meilleurs résultats obtenus sont ceux fondés sur les caractères uniquement, sans normalisation. L'utilisation d'un algorithme de clustering non-supervisé pour la création de vecteurs de mots (notamment `word2vec` (Mikolov *et al.*, 2013)) sur des gros volumes de données (notamment Wikipedia) a également été utilisé par l'INaLCO, Proxem (Marty *et al.*, 2015) et l'équipe IRISA (Vukotic *et al.*, 2015).

Enfin, les particularités inhérentes aux messages postés sur les réseaux sociaux en général (émoticônes, abréviations) et à Twitter en particulier (mots-dièses, noms d'utilisateur et présence de liens courts [3]), voir section 2, ont été prises en compte par plusieurs équipes (LIRMM ; LIMSI ; Synapse).

Lexiques Nous relevons que toutes les équipes ont utilisé des lexiques d'opinions, d'émotions et de sentiments, soit dans une chaîne de traitements (Synapse ; ANEX et LIDILEM par l'équipe mixte LINA/Dictanova), soit comme traits pour l'apprentissage statistique (CASOAR et EMOTAIX par l'équipe mixte IRIT/LIMSI, équipe 10 ; FEEL par le LIRMM ; Polarimots et DES – Dictionnaire Electronique des Synonymes par le LIMSI). Le LIMSI a également complété ces lexiques par des listes d'insultes tandis que l'équipe LINA/Dictanova (Hernandez *et al.*, 2015) a construit un lexique d'émoticônes composé de 40 classes.

5 Résultats

5.1 Évaluation officielle (par rapport à la référence)

5.1.1 Tâche 1

Sur cette tâche, les résultats en macro-précision (tableau 4) varient de 0 à 0,736, avec une moyenne de 0,582, une médiane de 0,693 et un écart-type de 0,238. L'analyse de la distribution des résultats montrent qu'un majorité des systèmes reconnaît la polarité des tweets avec une macro-précision comprise entre 0,54 et 0,75. Deux équipes (les équipes 4 et 25) ont cependant rencontré des problèmes techniques qui ne permettent pas l'obtention de résultats concluants.

3. La contrainte des 140 caractères maximum dans un message posté sur Twitter impose de réduire les URL. De nombreux services proposent ainsi des raccourcisseurs de liens internet tels que le service t.co propre à Twitter, dont le résultat sera de la forme `http://t.co/identifiant`.

Équipe	Soumissions			Classement
LIF (équipe 3)	**0,736**	0,722	0,688	1
INaLCO (équipe 2)	0,692	0,711	**0,734**	2
LIRMM (équipe 17)	0,732	0,725	**0,733**	3
Synapse (équipe 19)	**0,701**			4
Proxem (équipe 15)	**0,699**			5 ex-aequo
IRISA (équipe 14)	**0,699**	0,672	0,658	5 ex-aequo
LIMSI (équipe 23)	0,687	**0,688**		7
LINA / Dictanova (équipe 6)	0,655	**0,676**		8
IRIT / LIMSI (équipe 10)	0,577	0,578	**0,580**	9
TU Moldova (équipe 22)	**0,559**	0,547		10
LINA–Dimeco (équipe 25)	0,000	0,000	**0,136**	11
(équipe 4)	**0,041**			12

TABLE 4 – Résultats (macro-précision) par équipe sur la tâche 1, par résultats décroissants, la meilleure soumission de chaque équipe est en gras

5.1.2 Tâche 2.1

Sur cette tâche, les résultats en macro-précision (tableau 5) varient de 0,029 à 0,613, avec une moyenne de 0,514, une médiane de 0,217 et un écart-type de 0,029. Les résultats se répartissent selon deux grandes classes. Une première partie des systèmes catégorisent les tweets avec une macro-précision comprises entre 0,33 et 0,38, tandis qu'un deuxième ensemble de systèmes permettent de reconnaître les classes affectives génériques avec une macro-précision variant entre 0,5 et 0,62.

Équipe	Soumissions			Classement
LIRMM (équipe 17)	**0,613**	0,563	0,552	1
INaLCO (équipe 2)	0,572	0,562	**0,575**	2
IRISA (équipe 14)	**0,572**	0,478	0,502	3
LIF (équipe 3)	0,558	**0,560**	0,535	4
LINA / Dictanova (équipe 6)	0,508	**0,514**		5
TU Moldova (équipe 22)	**0,383**	0,382		6
IRIT / LIMSI (équipe 10)	0,269	**0,332**	**0,332**	7
LINA–Dimeco (équipe 25)	0,000	0,000	**0,097**	8
(équipe 4)	0,029	**0,029**		9

TABLE 5 – Résultats (macro-précision) par équipe sur la tâche 2.1, par résultats décroissants, la meilleure soumission de chaque équipe est en gras

5.1.3 Tâche 2.2

Sur cette tâche, les résultats en macro-précision (tableau 6) varient de 0 à 0,347, avec une moyenne de 0,180, une médiane de 0,200 et un écart-type de 0,152. Les résultats se répartissent selon trois grandes classes : un premier ensemble de systèmes réalisent une catégorisation fine des tweets avec une macro-précision inférieure à 0,05, la macro-précision des systèmes du deuxième groupe est comprise entre 0,17 et 0,23, tandis que le troisième ensemble de systèmes reconnaissent les catégories affectives fines avec une macro-précision variant entre 0,32 et 0,35.

5.2 Combinaison par vote pondéré (ROVER)

C'est John Fiscus (Fiscus, 1997) qui a proposé pour la première fois dans une campagne d'évaluation un algorithme de combinaison par vote pondéré des données produites par les participants. Cet algorithme, baptisé « ROVER » (*Reduced Output Voting Error Reduction*) par son auteur, a été créé pour une campagne d'évaluation sur la transcription automatique de parole organisée par le DARPA/NIST. Il permet à moindre coût d'augmenter la quantité de corpus annotés, de qualité et disponibles, en particulier réutilisables pour l'apprentissage automatique. Pour DEFT 2015 nous avons considéré une

Équipe	Soumissions			Classement
LIF (équipe 3)	**0,347**	0,327	0,327	1
INaLCO (équipe 2)	**0,337**	0,292	0,304	2
IRISA (équipe 14)	**0,325**	0,258	0,316	3
TU Moldova (équipe 22)	**0,226**	0,175		4
LIRMM (équipe 17)	0,037	**0,174**	0,007	5
LINA / Dictanova (équipe 6)	**0,028**	0,027		6
(équipe 4)	0,002	**0,002**		7
LINA–Dimeco (équipe 25)	0,000	0,000	0,000	8

TABLE 6 – Résultats (macro-précision) par équipe sur la tâche 2.2, par résultats décroissants, la meilleure soumission de chaque équipe est en gras

somme de vote pondérée par la mesure de performance obtenue par le participant qui a produit l'annotation. L'annotation retenue est celle qui obtient le score maximum. Les résultats de l'évaluation du ROVER appliqué aux tâches 1, 2.1 et 2.2 sont donnés dans le tableau 7. On observe que l'algorithme fonctionne bien s'il existe suffisamment de données pour chaque item d'annotation comme c'est le cas pour la tâche 1, alors que pour les autres tâches, il n'y a pas de gain de performance. Si l'on restreint la combinaison aux quatre systèmes les mieux placés (*ROVER4best* : INaLCO équipe 2, LIF équipe 3, IRISA équipe 14, et LIRMM équipe 17), on voit que le gain diminue un peu pour la tâche 1 mais reste positif tandis que la perte diminue pour les tâche 2.1 et tâche 2.2, confirmant l'intuition que moins l'on a de données, plus il faut sélectionner les données d'entrée en fonction de leur qualité (performance) pour ne retenir que les meilleures, l'effet de masse ne jouant plus pour éliminer le bruit.

tâche	max	ROVER	ROVER - max	ROVER4best	ROVER4best - max
tâche 1	0,736	**0,765**	**0,029**	**0,760**	**0.024**
tâche 2.1	0,613	0,589	-0,024	0,607	-0,006
tâche 2.2	0,347	0,330	-0,016	0,346	-0,001

TABLE 7 – Comparaison pour chaque tâche entre la meilleure performance des participants et celle du ROVER. max = meilleur résultat, ROVER = rover sur les meilleures soumissions de chaque système, ROVER - max = écart entre le rover et le meilleur système, ROVER4best = rover sur les meilleures soumissions des 4 meilleurs systèmes, ROVER4best - max = écart entre le rover4best et le meilleur système

6 Conclusion

Pour la troisième fois, le défi fouille de textes (DEFT) a proposé aux participants de travailler sur la fouille d'opinion. Contrairement aux éditions précédentes qui portaient sur des textes correctement rédigés (retranscriptions de débats parlementaires et articles de journaux) d'une part, et pour un nombre plutôt restreint de catégories d'autre part (favorable/défavorable, objectif/subjectif), l'édition 2015 s'est focalisée sur l'analyse complète des opinions, sentiments et émotions exprimées dans des messages postés sur le réseau social Twitter, sur la thématique du changement climatique. Cette analyse complète a été répartie en trois tâches : (i) déterminer la polarité globale des tweets, (ii) identifier les classes génériques (opinion, sentiment, émotion, information) et spécifiques (parmi 18 classes) de ces tweets, et (iii) analyser la source, la cible et l'expression porteuse d'opinion, de sentiment ou d'émotion.

Douze équipe ont participé. Les meilleurs résultats, en macro-précision, sont de 0,736 (polarité), 0,613 (classes génériques) et 0,347 (classes spécifiques). Les méthodes utilisées reposent majoritairement sur des approches par apprentissage statistique supervisé (SVM, Naïve Bayes, réseaux neuronaux, PPMC), et utilisent de nombreux lexiques d'opinions (ANEW, Casoar, Emotaix, Feel, Lidilem) et de polarités (Polarimots) comme traits. Il faut noter qu'aucun participant n'a soumis de données pour la dernière tâche, annotation fine des opinions, sentiment et émotions. Les premiers retours des participant indiquent qu'ils ont jugé la tâche trop difficile, les données d'entrainement ne leur permettant pas de construire une représentation exploitable.

Remerciements

Ce travail a été réalisé dans le cadre du projet uComp [4] financé par l'ERA Net CHIST-ERA (ANR-12-CHRI-0003).

A Définitions

Nous renseignons dans l'annexe suivante les définitions des 19 catégories fines du guide d'annotation (voir section 2.2).

A.1 Opinions

- *Accord* : opinion positive, la personne est d'accord avec au moins une autre personne sur un événement (*Tout à fait d'accord, le recyclage est devenu une nécessité*) ;
- *Désaccord* : opinion négative, la personne n'est pas d'accord (*Non aux éoliennes de la ferme du Torpt à Tourville et St - Meslin*) ;
- *Valorisation* : opinion positive, la personne désire une entité (événement, objet, personne) et a l'intention de réaliser une action en faveur de cette entité (*#tortueluth espèce menacée " Vu du ciel " #Gabon , les héros de la nature " super doc*) ;
- *Dévalorisation* : opinion négative, la personne ne désire pas une entité (événement, objet, personne) et n'a aucune intention de réaliser une action en faveur de cette entité (*Les saloperies promues par Royal : c' bruyant , laid , dévalorisant pour le foncier , et même pas efficace*) ;

A.2 Sentiments

- *Satisfaction* : sentiment positif, qui est suscité par la réalisation d'une intention résultant d'un désir (*Après un tel repas, je suis rassasié !*) ;
- *Insatisfaction* : sentiment négatif, qui est suscité par la non réalisation d'une intention résultant d'un désir (*Je n'ai pas pu partir en vélo*).

A.3 Emotions

- *Amour* : émotion positive, qui est suscité par le désir d'une autre personne ou animal (*L' amour et la fidelité sont des espéces en voie de disparition*) ;
- *Apaisement* : émotion positive, suscitée par la réalisation d'une intention suite à un événement non désiré (*Je suis soulagé, sa vie n'est plus en danger*) ;
- *Colère* : émotion négative, suscité par la réalisation d'un événement non désiré pare la personne et qui peut susciter ou pas une intention de réaction chez la personne (*Vent de colère sur nos villages . Éoliennes : l' arnaque totale*) ;
- *Déplaisir* : émotion négative qui résulte de la réalisation d'un événement non désiré par la personne (*Un lion en cage , un singe et deux serpents : une soirée cirque qui passe mal au Stamp #Waterloo*) ;
- *Dérangement* : émotion négative qui résulte de la réalisation d'un événement non désiré par la personne et qui suscite une intention d'action pour remédier à l'événement (*Les éoliennes vraiment source de nuisances*) ;
- *Ennui* : émotion négative, suscité par la connaissance de l'absence d'un événement désiré par la personne (*Je m'ennui, il y a rien d'intéressant à faire dans cette ville*) ;
- *Mépris* : émotion négative qui résulte d'une connaissance sur un objet, un événement, une personne qui est en opposition avec nos désirs (*Les losers et les saloppes , deux espèces en voie d' extinction*) ;
- *Peur* : émotion négative, suscité par la réalisation ou l'éventuelle réalisation d'un événement non désiré par la personne (*La sardine l' un des poissons les plus en danger en Méditerranée*) ;
- *Plaisir* : émotion positive, résulte de la réalisation d'un événement désiré par la personne (*je suis content que tu sois là*) ;
- *Surprise négative* : émotion négative, suscitée par la réalisation d'un événement non désiré et non attendu par la personne (*Mauvaise nouvelle : le ministre Henry a accordé le permis unique à Spe Luminus pour pour 5 éoliennes*) ;

4. http://www.ucomp.eu/

– *Surprise positive* : émotion positive, qui est suscitée par la réalisation d'un événement désirable et non attendu par la personne (*Bonne nouvelle pour Brest qui construira les jackets ! Saint - Brieuc .*) ;
– *Tristesse* : émotion négative, suscitée par la non réalisation d'un événement désiré et dont la réalisation est, soit possible dans le futur, soit impossible (*Je vois des trucs clignoter mais je sais pas si c' est des éoliennes ou des feux d' artifices c trist*).

A.4 Instruction

– *Instruction* : *Pensez à recycler vos , bouteilles vides ..*

Références

ABDAOUI A., TAPI NZALI M. D., AZÉ J., BRINGAY S., LAVERGNE C., MOLLEVI C. & PONCELET P. (2015). ADVANSE : Analyse du sentiment, de l'opinion et de l'émotion sur des tweets français. In *Actes de DEFT*, Caen, France : TALN.

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–96.

BOBICEV V. (2015). Sentiment detection using PPM. In *Actes de DEFT*, Caen, France : TALN.

CHARDON B., MULLER S., LAURENT D., PRADEL C. & SÉGUÉLA P. (2015). Chaîne de traitement symbolique pour l'analyse d'opinion – l'analyseur d'opinions de Synapse Développement face à Twitter. In *Actes de DEFT*, Caen, France : TALN.

FISCUS J. G. (1997). A post-processing system to yield reduced word error rates : recognizer output voting error reduction (rover). In *In proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, p. 347–357, Santa Barbara, CA.

FRAISSE A. & PAROUBEK P. (2014). Toward a unifying model for opinion, sentiment and emotion information extraction. In *Proc. of LREC*, Reykjavik, Iceland.

GROUIN C., ARNULPHY B., BERTHELIN J.-B., EL AYARI S., GARCIA-FERNANDEZ A., GRAPPY A., HURAULT-PLANTET M., PAROUBEK P., ROBBA I. & ZWEIGENBAUM P. (2009a). Présentation de l'édition 2009 du defi fouille de textes (deft'09). In *Actes de DEFT*, p. 35–50, Paris, France.

GROUIN C., BERTHELIN J.-B., EL AYARI S., HEITZ T., HURAULT-PLANTET M., JARDINO M., KHALIS Z. & LASTES M. (2007). Présentation de DEFT'07. In *Actes de DEFT*, Grenoble, France : AFIA.

GROUIN C., HURAULT-PLANTET M., PAROUBEK P. & BERTHELIN J.-B. (2009b). DEFT'07 : une campagne d'évaluation en fouille d'opinion. *Revue des Nouvelles Technologies de l'Information*, **RNTI E-17**, 1–24.

HERNANDEZ N., JADI G., LARK J. & MONCEAUX L. (2015). Exploitation de lexiques pour la catégorisation fine d'émotions, de sentiments et d'opinions. In *Actes de DEFT*, Caen, France : TALN.

LEJEUNE G. & DUMONCEAUX F. (2015). Une approche stylométrique pour la fouille d'opinion. In *Actes de DEFT*, Caen, France : TALN.

MANNING C. D. & SCHÜTZE H. (2000). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts : MIT Press.

MARTY J.-M., WENZEK G., SCHMITT E. & COULMANCE J. (2015). Analyse d'opinions de tweets par réseaux de neurones convolutionnels. In *Actes de DEFT*, Caen, France : TALN.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 746–751, Atlanta, Georgia : Association for Computational Linguistics.

MORLANE-HONDÈRE F. & D'HONDT E. (2015). Feature engineering for tweet polarity classification in the 2015 DEFT challenge. In *Actes de DEFT*, Caen, France : TALN.

VUKOTIC V., CLAVEAU V. & RAYMOND C. (2015). IRISA at DeFT 2015 : supervised and unsupervised methods in sentiment analysis. In *Actes de DEFT*, Caen, France : TALN.

WITTEN I. H. & FRANK E. (2005). *Data Mining - Pratical Machine Learning Tools and Techniques*. Morgan Kaufmann - Elsevier.