DEFT2018 : recherche d'information et analyse de sentiments dans des tweets concernant les transports en Île de France

Patrick Paroubek[1] Cyril Grouin[1] Patrice Bellot[2] Vincent Claveau[3]
Iris Eshkol-Taravella[4] Amel Fraisse[5] Agata Jackiewicz[6] Jihen Karoui[7]
Laura Monceaux[8] Juan-Manuel Torres-Moreno[9]
(1) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
(2) LSIS, Aix Marseille Université, CNRS, F-13397 Marseille
(3) CNRS, IRISA, Université de Rennes, F-35042 Rennes
(4) MoDyCo, Université Paris Nanterre, CNRS, F-92001 Nanterre
(5) GERIICO, Université de Lille, F-59653 Villeneuve-d'Ascq
(6) Praxiling, Université Paul Valéry Montpellier 3, CNRS, F-34199 Montpellier
(7) LIUM, Le Mans Université, F-72085 Le Mans
(8) LS2N, Université de Nantes, CNRS, Ecole centrale de Nantes, IMT Atlantique, F-44322 Nantes
(9) LIA, Université d'Avignon et des Pays de Vaucluse, F-84911 Avignon
{pap,grouin}@limsi.fr, patrice.bellot@lis-lab.fr,
vincent.claveau@irisa.fr, ieshkolt@parisnanterre.fr,
amel.fraisse@univ-lille3.fr, agata.jackiewicz@univ-montp3.fr,
jihen.karoui@univ-lemans.fr, Laura.Monceaux@univ-nantes.fr,
juan-manuel.torres@univ-avignon.fr

RÉSUMÉ

Cet article présente l'édition 2018 de la campagne d'évaluation DEFT (Défi Fouille de Textes). A partir d'un corpus de tweets, quatre tâches ont été proposées : identifier les tweets sur la thématique des transports, puis parmi ces derniers, identifier la polarité (négatif, neutre, positif, mixte), identifier les marqueurs de sentiment et la cible, et enfin, annoter complètement chaque tweet en source et cible des sentiments exprimés. Douze équipes ont participé, majoritairement sur les deux premières tâches. Sur l'identification de la thématique des transports, la micro F-mesure varie de 0,827 à 0,908. Sur l'identification de la polarité globale, la micro F-mesure varie de 0,381 à 0,823.

ABSTRACT

DEFT2018 : Information Retrieval and Sentiment Analysis in Tweets about Public Transportation in Île de France Region

This paper presents the 2018 DEFT text mining challenge. From a corpus of tweets, four tasks were proposed : first, to identify tweets about public transportation ; second, based on those tweets, to identify the global polarity (negative, neutral, positive, mixed), to identify clues of sentiment and target, and to annotate each tweet in terms of source and target concerning all expressed sentiments. Twelve teams participated, mainly on the two first tasks. On the identification of tweets about public transportation, micro F-measure values range from 0.827 to 0.908. On the identification of the global polarity, micro F-measure values range from 0.381 to 0.823.

MOTS-CLÉS : Classification automatique, Analyse de sentiments, Fouille de texte.

KEYWORDS: Automatic Classification, Sentiment Analysis, Text Mining.

1 Introduction

Dans la continuité de l'édition 2015 (Hamon *et al.*, 2015), la treizième édition du DÉfi Fouille de Textes (DEFT) porte sur l'extraction d'information et l'analyse de sentiments dans des tweets rédigés en français sur la thématique des transports. La campagne s'est déroulée sur une période limitée avec une ouverture des inscriptions le 31 janvier, la diffusion des données d'entraînement à partir du 19 février, et le déroulement de la phase de test entre le 28 mars et le 5 avril, sur une durée de trois jours fixée par chacun des participants. Quinze équipes se sont inscrites, dont une hors de France (Canada) et quatre issues d'entreprises privées. Au final, douze équipes auront participé (voir tableau 1).

Équipe	Nom de l'équipe	Affiliation	T1	T2	T3	T4
E1	EDF R&D	EDF R&D (entreprise)	X	X		
E2	CLaC	CLaC, Université Concordia	X	X		
E3	Tweetaneuse	STIH, Sorbonne Université ; LIPN, Université Paris 13	X	X		
E5	Synapse IRIT	Synapse Développement (entreprise) ; IRIT, Université Toulouse 3 Paul Sabatier	X	X		
E6	IRISA	IRISA, INSA Rennes	X	X	X	
E7	LIP6	LIP6, Sorbonne Université	X	X		
E8	Eloquant	Eloquant (entreprise)	X	X		
E9	EPITA	EPITA	X	X		
E10	UTTLM2S	Université Technologique de Troyes	X	X		
E11	Syllabs	Syllabs (entreprise)	X			
E14	ADVTeam	LIRMM, Université Montpellier, CNRS	X	X		
E15	LIS Lab	LIS, Aix Marseille Université		X		

TABLE 1 – Participation des équipes inscrites aux différentes tâches de la campagne DEFT2018

2 Corpus

Le corpus est constitué de tweets en français qui portent sur les transports en Île-de-France. Il contient 76 732 tweets sélectionnés parmi 80 000 tweets annotés manuellement. Les messages sont issus d'une sélection à base de mots-clés et de mesure d'entropie pour filtrer les doublons et les messages dépourvus de texte intelligible.

Chaque message a été annoté parmi cinq types de groupes et quatre types de relations.

— Groupes : la *SOURCE* est la séquence de mots faisant référence à l'entité qui exprime une subjectivité, la *CIBLE* est la séquence de mots faisant référence à l'entité sur laquelle porte cette subjectivité, l'*EXPRESSION d'OPINION/SENTIMENT/EMOTION (OSEE)* prend une valeur parmi les 18 catégories sémantiques polarisées proposés dans Fraisse & Paroubek (2014), le *MODIFIEUR* est marqueur d'intensité de l'expression de subjectivité et la *NEGATION* est un marqueur de négation

— Relations : *DIT* entre la source et l'OSEE, *SUR* entre l'OSEE et la cible, y compris les cibles intermédiaires, *MOD* entre le modifieur et l'OSEE, et *NEG* entre la négation et l'OSEE ou les cibles intermédiaires.

3 Description des tâches

3.1 Présentation

Autour de la thématique des transports, et sur la base des annotations précédentes, nous proposons quatre tâches, dont une très exploratoire sur l'analyse de sentiment à granularité sémantique fine.

Tâche 1 : classification transport/non-transport La première tâche vise à déterminer si un message concerne les transports ou non. Même si le message ne fait référence aux transports que de manière secondaire ou contextuelle, il sera considéré comme relatif aux transports.

— Transport : *Les gars qui puent des aisselles dans le bus c'est vous*
— Autre : *@InfoAbonneCanal bjr j' ai 2 décodeur canal , un me demande d' insérer la carte alors que dans le second les 2 carte fonctionne*

Tâche 2 : polarité globale La deuxième tâche consiste à déterminer la polarité globale d'un message concernant obligatoirement les transports, parmi quatre classes :

— POSITIF (message positif) : *J' ai trouvé une carte navigo dans le bus j' espère que la dame qui l' avait à Facebook sinon je vais pas pouvoir lui envoyer par la poste*
— NEGATIF (message négatif) : *Les gars qui puent des aisselles dans le bus c'est vous*
— NEUTRE (message factuel et objectif) : *Y' a une meuf elle a prit le bus pour s' arrêter à l' arrêt d' après , ils sont à 2 minutes l' un de l' autre à pied*
— MIXPOSNEG (message contenant des expressions positives et négatives, mais aucune des deux polarités ne domine) : *Bon voyage en mégabus malgré le retard le dimanche matin lyon marseille très bon bus manque un peu de confort le retour nikel le 7février*

Tâche 3 : marqueur de sentiment et cible Pour un message sur les transports exprimant des sentiments, cette tâche vise à déterminer, pour chaque expression du message : (i) l'empan de texte minimal qui renvoie à l'expression de sentiment, à l'exclusion des modifieurs et adjoints, et (ii) l'empan de texte maximal qui renvoie à la cible du sentiment, c'est-à-dire l'objet qu'il concerne, y compris modifieurs et adjoints.

Tâche 4 (exploratoire) : annotation complète Étant donné un message concernant les transports et exprimant des sentiments, la quatrième tâche vise à déterminer pour chaque expression de sentiment l'empan de texte minimal référant à l'expression de sentiment et les empans de texte maximaux référant respectivement à la CIBLE du sentiment (l'objet qu'il concerne) et à la SOURCE (l'entité qui exprime ce sentiment). Le cas échéant, on indiquera aussi les empans de texte minimaux en relation avec l'expression de sentiment qui référent à une cible ou un dérangement (voir figure 1).

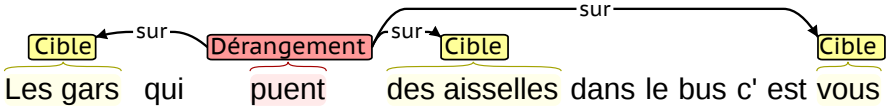

FIGURE 1 – Relations « sur » de l'expression de sentiment (catégorie « dérangement ») vers la cible

3.2 Statistiques

Nous présentons dans le tableau 2 de la distribution des 76 732 tweets annotés dans les corpus d'entrainement et de test, en fonction de la répartition entre catégories transport et non-transport utilisées dans la tâche 1 et de la polarité globale des tweets (POSITIF, NEGATIF, NEUTRE, MIXPOSNEG) utilisés dans la tâche 2 (seuls les tweets de la thématique des transports sont annotés en polarités).

	Transport		Autre		Total
	Entrainement	Test	Entrainement	Test	
POSITIF	7 328	857	*Pas d'annotation des polarités si autre thématique*		8 185
NEGATIF	13 109	1 525			14 634
NEUTRE	12 611	1 304			13 915
MIXPOSNEG	2 420	255			2 675
TOTAL	35 468	3 941	33 448	3 875	76 732

TABLE 2 – Distribution des tweets par corpus (entrainement, test), en fonction de la répartition entre catégories transport et non-transport, et de la polarité (POSITIF, NEGATIF, NEUTRE, MIXPOSNEG). Seuls les tweets de la thématique des transports sont annotés en polarités

4 Résultats

4.1 Mesures d'évaluation

Dans le cadre des deux premières tâches, nous utilisons les mesures habituelles de rappel, précision et F-mesure, calculées au moyen des micro-mesures (Manning & Schütze, 2000) : micro-rappel (formule 1, avec n le nombre total de classes) et micro-précision (formule 2, avec n le nombre total de classes), ainsi que la micro F-mesure calculée sur la base des résultats des deux précédentes mesures. Les micro-mesures attribuent un poids équivalent à chaque élément mesuré, indépendamment de la classe d'appartenance de cet élément. Les systèmes privilégiant les classes à fort effectif sont donc privilégiés dans ce mode de calcul, par opposition aux systèmes essayant de couvrir chacune des classes et notamment les classes à faible effectif.

$$\text{Micro-rappel} = \frac{\sum_{i=1}^{n} \text{vrais positifs}(i)}{\sum_{i=1}^{n} \text{vrais positifs}(i) + \sum_{i=1}^{n} \text{faux négatifs}(i)} \quad (1)$$

$$\text{Micro-précision} = \frac{\sum_{i=1}^{n} \text{vrais positifs}(i)}{\sum_{i=1}^{n} \text{vrais positifs}(i) + \sum_{i=1}^{n} \text{faux positifs}(i)} \quad (2)$$

4.2 Tâche 1 : classification transport/non-transport

Nous présentons dans le tableau 3 les décomptes en termes de vrais positifs, faux positifs et faux négatifs, ainsi que les valeurs de précision, rappel et F-mesure calculées au moyen des micro-mesures sur chacune des soumissions des participants à la tâche 1, classées par F-mesure décroissante. Nous indiquons le rang global de chaque équipe sur la base de la meilleure soumission. La figure 2 fournit

Rang global	Équipe et soumission	Décomptes			Micro-mesures		
		VP	FP	FN	Précision	Rappel	F-mesure
1	E3.R2 – Tweetaneuse	6497	1319	0	0,83124	1,00000	0,90785
2	E7.R4 – LIP6	6491	1325	0	0,83048	1,00000	0,90739
–	E7.R2 – LIP6	6490	1326	0	0,83035	1,00000	0,90731
–	E7.R5 – LIP6	6481	1335	0	0,82920	1,00000	0,90662
–	E7.R3 – LIP6	6478	1338	0	0,82881	1,00000	0,90639
3	E6.R2 – IRISA	6464	1352	0	0,82702	1,00000	0,90532
–	E6.R1 – IRISA	6461	1355	0	0,82664	1,00000	0,90509
–	E7.R1 – LIP6	6452	1364	0	0,82549	1,00000	0,90440
–	E6.R3 – IRISA	6449	1367	0	0,82510	1,00000	0,90417
–	E3.R3 – Tweetaneuse	6446	1369	1	0,82482	0,99984	0,90394
4	E5.R3 – Synapse IRIT	6443	1373	0	0,82433	1,00000	0,90371
5	E1.R1 – EDF R&D	6432	1384	0	0,82293	1,00000	0,90286
–	E5.R4 – Synapse IRIT	6425	1391	0	0,82203	1,00000	0,90232
–	E5.R1 – Synapse IRIT	6415	1401	0	0,82075	1,00000	0,90155
–	E6.R4 – IRISA	6414	1402	0	0,82062	1,00000	0,90148
–	E1.R2 – EDF R&D	6411	1405	0	0,82024	1,00000	0,90124
–	E5.R2 – Synapse IRIT	6399	1417	0	0,81871	1,00000	0,90032
6	E8.R2 – Eloquant	6362	1453	1	0,81408	0,99984	0,89745
7	E11.R3 – Syllabs	6300	1516	0	0,80604	1,00000	0,89260
–	E11.R2 – Syllabs	6299	1517	0	0,80591	1,00000	0,89253
8	E9.R2 – EPITA	6292	1524	0	0,80502	1,00000	0,89198
–	E3.R1 – Tweetaneuse	6289	1527	0	0,80463	1,00000	0,89174
–	E9.R3 – EPITA	6279	1537	0	0,80335	1,00000	0,89095
–	E9.R4 – EPITA	6279	1537	0	0,80335	1,00000	0,89095
–	E9.R1 – EPITA	6266	1550	0	0,80169	1,00000	0,88993
–	E11.R1 – Syllabs	6251	1565	0	0,79977	1,00000	0,88875
–	E11.R4 – Syllabs	6243	1573	0	0,79875	1,00000	0,88811
–	E9.R5 – EPITA	6238	1578	0	0,79811	1,00000	0,88772
9	E10.R1 – UTTLM2S	6220	1596	0	0,79580	1,00000	0,88629
–	E8.R1 – Eloquant	6202	1613	1	0,79360	0,99984	0,88486
–	E11.R5 – Syllabs	6191	1625	0	0,79209	1,00000	0,88399
–	E5.R5 – Synapse IRIT	6187	1629	0	0,79158	1,00000	0,88367
10	E2.R2 – CLaC	6093	1723	0	0,77955	1,00000	0,87612
–	E10.R2 – UTTLM2S	6067	1749	0	0,77623	1,00000	0,87402
–	E3.R4 – Tweetaneuse	6046	1769	1	0,77364	0,99983	0,87231
–	E10.R3 – UTTLM2S	6015	1801	0	0,76958	1,00000	0,86979
11	E14.R1 – ADVTeam	5511	2305	0	0,70509	1,00000	0,82704
–	E2.R1 – CLaC	4387	3428	1	0,56136	0,99977	0,71900

TABLE 3 – Décomptes de vrais positifs (VP), faux positifs (FP) et faux négatifs (FN), et résultats (micro-mesures) sur la tâche 1 (classification transport/autre) classés par F-mesure décroissante

une représentation tri-dimensionnelle des valeurs de précision, rappel et F-mesure pour chaque soumission de chaque équipe sur la première tâche. La meilleure valeur de F-mesure est mise en évidence par une flèche (équipe 3, deuxième soumission).

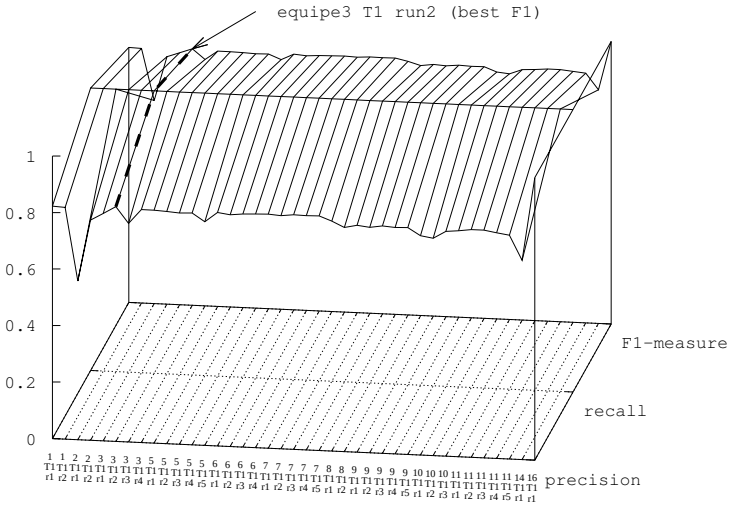

FIGURE 2 – Représentation tri-dimensionnelle des valeurs de précision, rappel et F-mesure pour chaque soumission de chaque équipe sur la première tâche. Le dernier résultat à droite représente la performance des données de référence

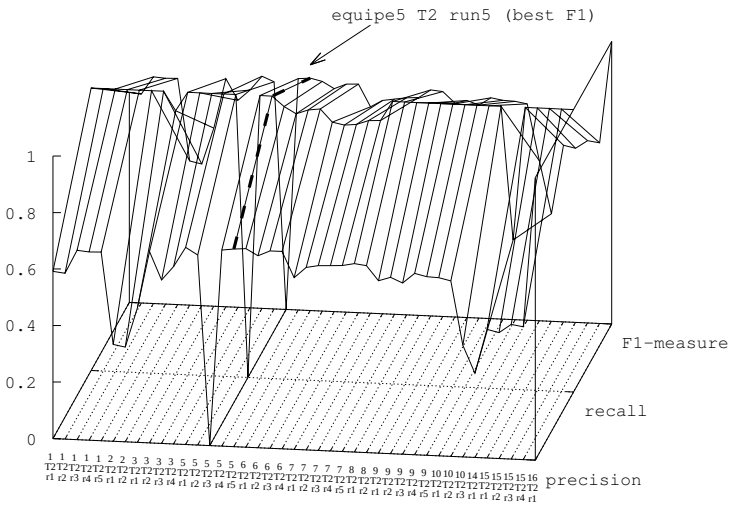

FIGURE 3 – Représentation tri-dimensionnelle des valeurs de précision, rappel et F-mesure pour chaque soumission de chaque équipe sur la deuxième tâche. Le dernier résultat à droite représente la performance des données de référence

4.3 Tâche 2 : polarité globale

Nous reportons dans le tableau 4 les décomptes en termes de vrais positifs, faux positifs et faux négatifs, ainsi que les valeurs de précision, rappel et F-mesure calculées au moyen des micro-mesures sur chacune des soumissions des participants à la tâche 2, classées par F-mesure décroissante. Nous indiquons le rang global de chaque équipe sur la base de la meilleure soumission. La figure 3 fournit une représentation tri-dimensionnelle des valeurs de précision, rappel et F-mesure pour chaque soumission de chaque équipe sur la deuxième tâche. La meilleure valeur de F-mesure est mise en évidence par une flèche (équipe 5, cinquième soumission).

4.4 Significativité statistique

Nous observons une grande homogénéité des résultats, plusieurs équipes obtenant des résultats proches sur plusieurs soumissions. Nous avons alors vérifié les différences entre résultats en mesurant la significativité statistique des différences de résultats. Pour cela, nous avons mis en place le protocole suivant : le jeu de test a été découpé aléatoirement en 30 morceaux et les performances des systèmes sur ces 30 morceaux (valeurs de F-mesure calculées par rapport à la référence) ont été comparées à l'aide d'un t-test pairé et d'un test de Wilcoxon. Le seuil de significativité est fixé à 5 % (p-valeur=0.05).

Les résultats des tests de significativité statistique (t-test) calculés entre soumissions prises deux à deux sont représentés dans les tableaux 5 (première tâche) et 6 (deuxième tâche). Les lignes de ces deux tableaux sont classées par valeurs de micro F-mesure décroissantes, alors que les colonnes sont classées par ordre d'inscription. Ces tableaux se lisent comme suit. La meilleure soumission sur la première tâche (première ligne du tableau 5) est E3.R2 (run 2 de l'équipe Tweetaneuse sachant que E3 correspond à l'équipe Tweetaneuse, cf. tableau 1). Dans la colonne correspondante, on observe des 'o' sur les cinq soumissions suivantes (de E7.R4 à E6.R2), ce qui signifie que les différences observées entre la soumission E3.R2 et les cinq suivantes ne sont pas significatives (p-valeur<0,05). Les différences deviennent significatives (symbole '*') à partir de la soumission E6.R1 (au 7ème

Rang global	Équipe et soumission	Décomptes			Micro-mesures		
		VP	FP	FN	Précision	Rappel	F-mesure
1	E5.R5 – Synapse IRIT	2755	1186	0	0,69906	1,00000	0,82288
–	E5.R1 – Synapse IRIT	2748	1193	0	0,69728	1,00000	0,82165
–	E5.R4 – Synapse IRIT	2734	1207	0	0,69373	1,00000	0,81918
2	E6.R1 – IRISA	2700	1143	98	0,70258	0,96497	0,81313
–	E6.R4 – IRISA	2678	1158	105	0,69812	0,96227	0,80919
3	E3.R2 – Tweetaneuse	2668	1273	0	0,67699	1,00000	0,80738
–	E6.R3 – IRISA	2667	1153	121	0,69817	0,95660	0,80720
–	E5.R2 – Synapse IRIT	2650	1291	0	0,67242	1,00000	0,80413
4	E1.R3 – EDF R&D	2641	1300	0	0,67013	1,00000	0,80249
–	E1.R5 – EDF R&D	2631	1310	0	0,66760	1,00000	0,80067
5	E8.R1 – Eloquant	2628	1313	0	0,66684	1,00000	0,80012
–	E1.R4 – EDF R&D	2625	1316	0	0,66607	1,00000	0,79957
–	E8.R2 – Eloquant	2607	1334	0	0,66151	1,00000	0,79627
–	E6.R2 – IRISA	2569	1212	160	0,67945	0,94137	0,78925
6	E9.R4 – EPITA	2529	1412	0	0,64172	1,00000	0,78176
–	E9.R5 – EPITA	2491	1450	0	0,63207	1,00000	0,77456
7	E10.R1 – UTTLM2S	2483	1458	0	0,63004	1,00000	0,77304
8	E7.R5 – LIP6	2473	1284	184	0,65824	0,93075	0,77113
–	E3.R4 – Tweetaneuse	2473	1467	1	0,62766	0,99960	0,77113
–	E9.R2 – EPITA	2459	1482	0	0,62395	1,00000	0,76844
–	E7.R4 – LIP6	2459	1292	190	0,65556	0,92827	0,76844
–	E10.R2 – UTTLM2S	2446	1495	0	0,62065	1,00000	0,76593
–	E7.R3 – LIP6	2426	1284	231	0,65391	0,91306	0,76205
–	E9.R1 – EPITA	2404	1537	0	0,61000	1,00000	0,75776
–	E7.R2 – LIP6	2390	1311	240	0,64577	0,90875	0,75502
–	E9.R3 – EPITA	2386	1555	0	0,60543	1,00000	0,75423
–	E1.R1 – EDF R&D	2332	1609	0	0,59173	1,00000	0,74350
–	E1.R2 – EDF R&D	2313	1628	0	0,58691	1,00000	0,73969
–	E3.R3 – Tweetaneuse	2279	1662	0	0,57828	1,00000	0,73280
–	E7.R1 – LIP6	2269	1468	204	0,60716	0,91751	0,73076
9	E15.R3 – LIS Lab	1877	2064	0	0,47628	1,00000	0,64524
–	E15.R4 – LIS Lab	1852	2089	0	0,46993	1,00000	0,63939
–	E3.R1 – Tweetaneuse	1814	1997	130	0,47599	0,93313	0,63041
–	E15.R1 – LIS Lab	1793	2148	0	0,45496	1,00000	0,62539
–	E15.R2 – LIS Lab	1755	2186	0	0,44532	1,00000	0,61622
–	E10.R3 – UTTLM2S	1556	2385	0	0,39482	1,00000	0,56613
10	E2.R1 – CLaC	1350	2591	0	0,34255	1,00000	0,51030
–	E2.R2 – CLaC	1320	2621	0	0,33494	1,00000	0,50181
11	E14.R1 – ADVTeam	927	2186	828	0,29778	0,52821	0,38085
–	E5.R3 – Synapse IRIT	0	0	3941	0,00000	0,00000	0,00000

TABLE 4 – Décomptes de vrais positifs (VP), faux positifs (FP) et faux négatifs (FN), et résultats (micro-mesures) sur la tâche 2 (polarité globale) classés par F-mesure décroissante

rang global). De même, les colonnes des soumissions E7.R4 à E6.R2 permettent de constater que les différences ne sont pas significatives. Ainsi émergent des clusters de soumissions, avec un premier cluster composé des runs 2 de Tweetaneuse (E3), des runs 4, 2, 5, 3 du LIP6 (E7) et du run 2 de l'IRISA (E6). Ce phénomène est moins marqué sur la deuxième tâche. Nous observons que seuls les trois premiers runs ne présentent aucune différence du point de vue de nos tests statistiques.

	E1.R1	E1.R2	E3.R1	E3.R2	E3.R3	E3.R4	E5.R1	E5.R2	E5.R3	E5.R4	E5.R5	E6.R1	E6.R2	E6.R3	E6.R4	E7.R1	E7.R2	E7.R3	E7.R4	E7.R5	E8.R1	E8.R2	E9.R1	E9.R2	E9.R3	E9.R4	E9.R5	E11.R1	E11.R2	E11.R3	E11.R4	E11.R5
E3.R2	*	*	*		*	*	*	*	*	*	*	*	o	*	*	*	o	o	o	o	*	*	*	*	*	*	*	*	*	*	*	*
E7.R4	*	*	*	o	*	*	*	*	*	*	*	o	o	*	*	*	o	o		o	*	*	*	*	*	*	*	*	*	*	*	*
E7.R2	*	*	*	o	*	*	*	*	*	*	*	*	o	*	*	*		o	o	o	*	*	*	*	*	*	*	*	*	*	*	*
E7.R5	*	*	*	o	*	*	*	*	o	*	*	o	o	o	*	*	o	o	o		*	*	*	*	*	*	*	*	*	*	*	*
E7.R3	*	*	*	o	*	*	*	*	o	*	*	o	o	o	*	o	o		o	o	*	*	*	*	*	*	*	*	*	*	*	*
E6.R2	*	*	*	o	*	*	*	*	o	*	*	o		o	*	o	o	o	o	o	*	*	*	*	*	*	*	*	*	*	*	*
E6.R1	*	*	*	*	*	*	*	*	o	*	*		o	o	*	o	*	o	o	o	*	*	*	*	*	*	*	*	*	*	*	*
E7.R1	o	o	*	*	*	*	*	*	o	o	*	o	o	o	*		*	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E6.R3	o	*	*	*	*	*	o	*	o	o	*	o	o		*	o	*	o	*	o	*	*	*	*	*	*	*	*	*	*	*	*
E3.R3	o	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	o	*	o	*	o	*	*	o	*
E5.R3	o	*	*	*	*	*	*	*		*	*	o	o	o	*	o	*	o	*	o	*	*	*	*	*	*	*	*	*	*	*	*
E1.R1		o	o	*	o	o	o	o	o	o	o	*	*	o	o	o	*	*	*	*	o	o	o	o	o	o	o	o	o	o	o	o
E5.R4	o	o	*	*	*	*	o	o	*		*	*	*	o	o	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E5.R1	o	o	*	*	*	*		o	*	o	*	*	*	o	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E6.R4	o	o	*	*	*	*	o	o	*	o	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E1.R2	o		*	*	*	*	o	o	*	o	*	*	*	*	o	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E5.R2	o	o	*	*	*	*	o		*	o	*	*	*	*	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E8.R2	o	*	*	*	*	*	*	*	*	*	o	*	*	*	*	*	*	*	*	*	*		*	o	*	o	*	o	*	*	o	o
E11.R3	o	*	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	o	*	o	o	o	*	o		*	*
E11.R2	o	*	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	o	*	o	o	o	*		o	*	*
E9.R2	o	*	o	*	o	o	*	*	*	*	o	*	*	*	*	*	*	*	*	*	*	o	*		*	o	o	o	*	*	o	o
E3.R1	o	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	o	o	o	o	o	o	o	o	o	o	*
E9.R3	o	*	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	o	*		o	o	o	o	o	o	o	*
E9.R4	o	*	o	*	o	o	*	*	*	*	o	*	*	*	*	*	*	*	*	*	*	o	o	o		o	o	o	o	o	o	*
E9.R1	o	*	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		*	o	o	o	o	o	o	o	*
E11.R1	o	*	o	*	o	o	*	*	*	*	o	*	*	*	*	*	*	*	*	*	*	o	o	o	o	o	o		*	*	o	*
E11.R4	o	*	o	*	o	o	*	*	*	*	o	*	*	*	*	*	*	*	*	*	*	o	o	o	o	o	o	o		*		*
E9.R5	o	*	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	o	o	o	o		o	o	o	o	*
E8.R1	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*
E11.R5	o	*	*	*	*	*	*	*	*	*	o	*	*	*	*	*	*	*	*	*	*	o	*	o	*	*	*	*	*	*	*	
E5.R5	o	*	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	o	*	o	*	o	*	o	*	*	o	o
E3.R4	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	o	*	o	*	o	*	*	o	*

TABLE 5 – Différence entre résultats significative (symbole '*', p-value $< 0{,}05$ sur le t-test) et non significative (symbole 'o') sur la tâche 1, dans l'ordre de classement (cf. tableau 3)

	E1.R1	E1.R2	E1.R3	E1.R4	E1.R5	E5.R1	E5.R2	E5.R4	E5.R5	E7.R1	E7.R2	E7.R3	E7.R4	E7.R5	E8.R1	E8.R2	E15.R1	E15.R2	E15.R3	E15.R4
E5.R5	*	*	*	*	*	o	o	o		*	*	*	*	*	*	*	*	*	*	*
E5.R1	*	*	*	*	*		o	o	o	*	*	*	*	*	*	*	*	*	*	*
E5.R4	*	*	*	*	*	o	*		o	*	*	*	*	*	*	*	*	*	*	*
E5.R2	o	o	o	o	o	o		*	o	o	o	o	o	o	o	o	o	o	o	o
E1.R3	o	o		o	o	*	o	*	*	o	o	o	o	o	o	o	o	o	o	o
E1.R5	*	*	o	o		*	o	*	*	*	*	*	*	*	o	o	*	*	*	*
E8.R1	o	o	o	o	o	*	o	*	*	o	o	o	o	o		o	o	o	o	o
E1.R4	*	*	o		o	*	o	*	*	*	*	*	*	*	o	o	*	*	*	*
E8.R2	*	*	o	o	o	*	o	*	*	*	*	*	*	*	o		*	*	*	*
E7.R5	*	*	o	*	*	*	o	*	*	*	*	*	o		o	*	*	*	*	*
E7.R4	*	*	o	*	*	*	o	*	*	*	*	o		o	o	*	*	*	*	*
E7.R3	*	*	o	*	*	*	o	*	*	*	o		o	*	o	*	*	*	*	*
E7.R2	o	*	o	*	*	*	o	*	*	*		o	*	*	o	*	*	*	*	*
E1.R1		o	o	*	*	*	o	*	*	*	o	*	*	*	o	*	*	*	*	*
E1.R2	o		o	*	*	*	o	*	*	o	*	*	*	*	o	*	o	o	o	o
E7.R1	*	o	o	*	*	*	o	*	*		*	*	*	*	o	*	o	o	o	o
E15.R3	*	o	o	*	*	*	o	*	*	o	*	*	*	*	o	*	*	*		*
E15.R4	*	o	o	*	*	*	o	*	*	o	*	*	*	*	o	*	*	*	*	
E15.R1	*	o	o	*	*	*	o	*	*	o	*	*	*	*	o	*		*	*	*
E15.R2	*	o	o	*	*	*	o	*	*	o	*	*	*	*	o	*	*		*	*

TABLE 6 – Différence entre résultats significative (symbole '*', p-value $< 0{,}05$ sur le t-test) et non significative (symbole 'o') sur la tâche 2, dans l'ordre de classement (cf. tableau 4)

4.5 Méthodes des participants

L'utilisation actuelle de méthodes par apprentissage statistique dans les différentes tâches de TAL et la disponibilité de données annotées dans ce défi ont conduit la majorité des participants à se tourner vers des méthodes d'apprentissage supervisé. Les réseaux de neurones convolutifs (CNN) et récurrents (LSTM, biLSTM, et GRU), complétés par des plongements lexicaux, ont ainsi largement été utilisés, tant par les participants d'entreprises tels que EDF R&D (Suignard *et al.*, 2018) ou Synapse Développement & IRIT (Sileo *et al.*, 2018), que par les participants académiques avec les équipes CLaC (Jacques *et al.*, 2018), de l'EPITA (Sainson *et al.*, 2018), de l'IRISA (Minard *et al.*, 2018), du LIP6 (Dias *et al.*, 2018), du LIRMM (Azmy *et al.*, 2018), du LIS (Htait, 2018) et Tweetaneuse (Buscaldi *et al.*, 2018). Les algorithmes d'apprentissage supervisé traditionnels (arbres de décision, bayésien naïf, entropie maximale, CRF, SVM) ont été plus rarement utilisés, parfois à titre de comparaison avec d'autres méthodes, par l'entreprise Syllabs (Monnin *et al.*, 2018) et par les équipes CLaC (Jacques *et al.*, 2018), IRISA (Minard *et al.*, 2018) et Tweetaneuse (Buscaldi *et al.*, 2018). L'entreprise Eloquant (Graceffa *et al.*, 2018) a employé des méthodes symboliques, en adaptant aux spécificités du défi la méthode employée en interne pour traiter des données de relations clients. Cette adaptation passe notamment par un enrichissement sémantique et une prise en compte des propriétés de surface des messages postés sur les réseaux sociaux.

5 Conclusion

L'édition 2018 du défi fouille de texte (DEFT) s'est révélée un succès en terme de nombre de participants, portée par une thématique connue mais toujours en vogue (la fouille d'opinion) et un ensemble de techniques (classification, notamment par réseaux de neurones récurrents) faciles à mettre-en-œuvre. Parmi les quatre tâches proposées, seules les deux premières ont été traitées par les participants. Sur l'identification de la thématique des transports, la micro F-mesure varie de 0,827 à 0,908 tandis que pour l'identification de la polarité globale, la micro F-mesure varie de 0,381 à 0,823.

Nous observons des résultats très homogènes entre participants sur chacune des deux premières tâches, que nous pouvons expliquer par plusieurs points :

— les techniques utilisées sont similaires ;
— la première tâche était relativement facile, probablement parce que la sélection initiale des tweets s'est faite sur mots-clés, amenant les systèmes à des performances quasi optimales ;
— la deuxième tâche, plus difficile au vu des résultats qui doivent être relativisés par le bruit résiduel présent dans les annotations, a aussi amené les systèmes à une sorte de plafond, difficile à dépasser au regard des données.

Le faible nombre de participants (une seule équipe) sur la tâche 3 et l'absence de participants sur la tâche exploratoire est un peu décevant. Bien que nécessitant des techniques d'annotations de texte plutôt que de classification comme sur les deux premières tâches, de nombreux outils et retours d'expérience sur des tâches similaires sont disponibles et auraient pu permettre à un plus grand nombre de participer. Le calendrier serré de cette édition (au maximum un mois et demi pour la phase d'entraînement) aura pu décourager certaines équipes de s'engager dans ces tâches plus complexes.

Remerciements

Le corpus de la campagne d'évaluation DEFT2018 a été produit dans le cadre du projet REQUEST (Programme d'Investissement d'Avenir, appel Cloud computing & Big Data, convention 018062-25005) et annoté en collaboration avec ELDA. Le projet MIROR, du programme de recherche et d'innovation de l'Union Européenne Horizon 2020, « Marie Sklodowska-Curie grant agreement No 676207 », a contribué à l'organisation de la campagne DEFT 2018.

Références

AZMY W. M., MOULAHI B., BRINGAY S., AZÉ J. & SERVAJEAN M. (2018). Lirmm@deft-2018 – modèle de classification de la vectorisation des documents. In *Actes de DEFT*, Rennes, France.

BUSCALDI D., LE ROUX J. & LEJEUNE G. (2018). Modèles en caractères pour la détection de polarité dans les tweets. In *Actes de DEFT*, Rennes, France.

DIAS C.-E., GAINON DE FORSAN DE GABRIAC C., GUIGUE V. & GALLINARI P. (2018). DEFT 2018 : Attention sélective pour classification de microblogs. In *Actes de DEFT*, Rennes, France.

FRAISSE A. & PAROUBEK P. (2014). Toward a unifying model for opinion, sentiment and emotion information extraction. In *Proc of LREC*, p. 3881–3886, Reykjavik, Iceland.

GRACEFFA D., RAMOND A., DUSSERRE E., KALITVIANSKI R., RUHLMANN M. & PADRÓ M. (2018). Notre tweet première fois au DEFT-2018 : systèmes de détection de polarité et de transports. In *Actes de DEFT*, Rennes, France.

HAMON T., FRAISSE A., PAROUBEK P., ZWEIGENBAUM P. & GROUIN C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (DEFT). In *Actes de DEFT*, Caen, France.

HTAIT A. (2018). Adapted sentiment similarity seed words for french tweets' polarity classification. In *Actes de DEFT*, Rennes, France.

JACQUES S., FARAHNAK F. & KOSSEIM L. (2018). CLaC @ DEFT 2018 : Sentiment analysis of tweets on transport from île-de-France. In *Actes de DEFT*, Rennes, France.

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

MINARD A.-L., RAYMOND C. & CLAVEAU V. (2018). Participation de l'IRISA à DeFT 2018 : classification et annotation d'opinion dans des tweets. In *Actes de DEFT*, Rennes, France.

MONNIN C., QUERNÉ O. & HAMON O. (2018). Syllabs@DEFT2018 : combinaison de méthodes de classification supervisées. In *Actes de DEFT*, Rennes, France.

SAINSON A., LINSENMAIER H., MAJED A., CADET X. & BOUCHEKIF A. (2018). LSE au DEFT 2018 : Classification de tweets basée sur les réseaux de neurones profonds. In *Actes de DEFT*, Rennes, France.

SILEO D., VAN DE CRUYS T., MULLER P. & PRADEL C. (2018). Concaténation de réseaux de neurones pour la classification de tweets, DEFT2018. In *Actes de DEFT*, Rennes, France.

SUIGNARD P., CHARAUDEAU L., BOUMGHAR M., BOTHUA M. & LAGARDE D. (2018). Participation d'EDF R&D à DEFT 2018. In *Actes de DEFT*, Rennes, France.