Notre tweet première fois au DEFT-2018 : systèmes de détection de polarité et de transports

David Graceffa Armelle Ramond Emmanuelle Dusserre
Ruslan Kalitvianski Mathieu Ruhlmann Muntsa Padró
Eloquant, 5 allée de Palestine, 38610 Gières, France
{prenom.nom}@eloquant.com

RESUME

Cet article décrit les systèmes de l'équipe Eloquant pour la catégorisation de tweets en français dans les tâches 1 (détection de la thématique *transports en commun*) et 2 (détection de la polarité globale) du DEFT 2018. Nos systèmes reposent sur un enrichissement sémantique, l'apprentissage automatique et, pour la tâche 1 une approche symbolique. Nous avons effectué deux *runs* pour chacune des tâches. Nos meilleures F-mesures (0.897 pour la tâche 1 et 0.800 pour la tâche 2) sont au-dessus de la moyenne globale pour chaque tâche, et nous placent dans les 30% supérieurs de tous les *runs* pour la tâche 2.

ABSTRACT

Systems for detecting polarity and public transport discussions in French tweets
This paper presents Eloquant's team's systems for automatic classification of French tweets in tasks 1 (detection of discussions about public transport) and 2 (detection of the overall polarity) of DEFT-2018. Our systems are based on semantic enrichment, machine learning and, for task 1, a symbolic approach. We performed two runs for each task. Our best F-measures (0.897 for task 1 and 0.800 for task 2) are above the overall average for each task and place us in the top 30% of all runs for task 2.

MOTS-CLES : Tweets, fouille d'opinions, transports, classification automatique
KEYWORDS : Tweets, sentiment analysis, transport, automatic classification

1 Introduction

Cet article décrit les méthodes et les résultats obtenus par l'équipe sémantique d'Eloquant aux tâches 1 et 2 du *Défi Fouille de Textes 2018* (Paroubek et al., 2018), qui concerne l'analyse des thématiques et des opinions exprimées dans un corpus de 68 916 tweets annotés manuellement[1] dans le cadre du projet REQUEST[2].

La **tâche 1** consiste à déterminer automatiquement si un tweet concerne ou non les transports en commun. Il s'agit donc de classification binaire. La **tâche 2** est également une tâche de classification

[1] https://ocsync.limsi.fr/index.php/s/Mbm4Hl5YnALJRKx

[2] Programme d'Investissement d'Avenir, appel Cloud computing & Big Data, convention 018062-25005

et consiste à déterminer, pour les tweets qui concernent les transports en commun (51% du total), leur polarité globale parmi quatre catégories : *positif*, *négatif*, *neutre* et *mixposneg* (à la fois positif et négatif). La partie du corpus d'entrainement annotée par polarité contient environ 37% de tweets négatifs, 36% de tweets neutres, 21% de tweets positifs et 7% de tweets à polarité mixte.

Les outils d'analyse sémantique d'Eloquant ont été développés dans le but de traiter des *verbatim* issus du domaine de la relation client (SMS ou formulaires web). Il n'était donc *a priori* pas trivial pour nous de réaliser les deux tâches (1 et 2) avec nos méthodes symboliques existantes car elles sont composées de règles très spécifiques au style de notre secteur. En effet, l'attitude langagière adoptée dans le cadre de la relation client et celle que l'on trouve sur Twitter sont extrêmement différentes. C'est pourquoi, nous avons suivi deux voies : développer pour l'occasion une méthode symbolique (tâche 1, *run* 1), et proposer des méthodes entièrement statistiques (tâche 1 *run* 2, et tâche 2 *run* 1 et 2). Ces deux approches reposent sur un enrichissement sémantique, nous en parlerons plus précisément dans la suite de cet article.

Dans le reste de l'article nous décrivons les étapes (communes pour toutes les tâches et spécifiques pour chaque tâche) de traitement du corpus, de création de classifieurs, leurs évaluations et la mise en perspective des résultats par rapport à la qualité des données et les limites théoriques des approches utilisées.

2 Chaîne de traitement des données

Qu'il s'agisse de la phase d'apprentissage d'une méthode statistique, de la phase de test ou encore d'une méthode symbolique, nous avons procédé à un travail de représentation des documents, qui se traduit par un pré-traitement, une analyse morphosyntaxique et enfin un enrichissement sémantique.

2.1 Pré-traitement et analyse morphosyntaxique

Afin de préparer notre corpus aux différentes méthodes testées, nous avons effectué une suite de normalisations. En effet, les données issues de Twitter étant particulièrement bruitées il nous est apparu primordial d'en homogénéiser le contenu, en plus de nos traitements habituels.
Nous avons appliqué nos scripts génériques de correction d'encodage des caractères accentués pour les ramener à de l'UTF-8. Ainsi, la séquence "ã©" a par exemple été corrigée en "é" ou encore """ par le caractère ". Les chaînes "[ASCII012CTRLC]" et "[ASCII015CTRLC]" rencontrées dans certains tweets, et qui correspondent à des passages à la ligne, ont été remplacées par un espace lorsqu'ils étaient précédés par un caractère de ponctuation, sinon par un point.

Nous avons également normalisé les émojis en ajoutant un espace lorsqu'apparaissaient plusieurs emojis collés, afin qu'ils puissent être reconnus comme des *tokens*[3] distincts, par exemple ☺☺☺ devient ☺ ☺ ☺. Cela est nécessaire pour l'enrichissement sémantique qui intervient par la suite.

[3] Dans cet article nous préférons cet anglicisme au terme français « item » moins courant, ou « forme », plus ambigu.

Ensuite, nous avons procédé à une normalisation orthographique :

- Suppression des espaces superflus.
- Reponctuation par un point des phrases non ponctuées.
- Suppression des URL.
- Correction orthographique à l'aide de ressources constituées habituellement pour notre cœur de métier, la relation client.
- Suppression de lettres dupliquées dans certaines formes telles que "xptdrrr", "mdrrr", "ptdrrr", "loool". (Ces formes peuvent être des marqueurs importants pour la détection de polarité.)

Enfin, nous avons appliqué les analyses et traitements classiques : segmentation en phrases, tokenisation, lemmatisation, attribution des étiquettes morphosyntaxiques[4], dépendances syntaxiques.

2.2 Enrichissement sémantique

La littérature (Abdaoui et al., 2015, Chen et al., 2011, Vernier et al., 2009) indique que l'enrichissement sémantique peut améliorer les résultats dans des tâches de classification. Nous avons mis cela en œuvre grâce à des *gazetteers* : il s'agit de listes de lemmes où chaque entrée est associée à un *tag* sémantique. Chaque *gazetteer* a été construit et appliqué en fonction d'un type de *PoS* afin de réduire l'ambiguïté : lorsque le système rencontre dans un document un *token* présent dans un *gazetteer* avec la *PoS* correspondante, il attribue à ce *token* le *tag* sémantique associé dans le *gazetteer*. L'annotation sémantique qui en découle peut ensuite être exploitée comme trait dans les règles symboliques ou dans l'apprentissage automatique.

Pour la tâche 1, nous considérons qu'une annotation du lexique relevant de la thématique des transports en commun est pertinente. Nous avons réalisé une extraction semi-automatique des lemmes spécifiques à ce domaine, en plusieurs étapes : extraction des concepts selon la méthode proposée par (Sclano et Velardi, 2007), calcul pour chaque concept de deux fréquences relatives (dans le corpus *transport* et dans le corpus non *transport*), filtrage sur les concepts ayant une fréquence relative supérieure à 0,005 dans l'un ou l'autre des corpus, tri des concepts par ordre décroissant de la différence entre les deux fréquences. La liste ainsi obtenue a ensuite été revue et complétée manuellement afin de créer les *gazetteers* en attribuant à chaque entrée le *tag* sémantique TRANSPORT.

Pour la tâche 2, nous avons également relevé du lexique porteur de polarité (positive ou négative), de façon manuelle cette fois en lisant un certain nombre de tweets. En effet, notre méthode d'extraction semi-automatique décrite plus haut n'est pas adaptée à des notions subjectives comme la polarité. A l'issue de ce travail, nous disposons de *gazetteers* dans lesquels les lemmes relevés sont associés aux *tags* POS[5] ou NEG selon leur sémantique. De plus, l'exercice est un peu plus complexe car il faut ensuite prendre en compte une éventuelle négation (par exemple "Je ne suis pas content"), nous avons donc ajouté une couche de règles symboliques pour traiter les principaux cas rencontrés. Ainsi, l'annotation POS attribuée par un *gazetteer* au *token* "content" va être transformée en NEG par l'application d'une règle de détection de la négation.

[4] Par souci de brièveté nous utiliserons l'abréviation anglaise "*PoS*" par la suite.

[5] POS pour positif, à ne pas confondre avec *PoS*.

3 Création des classifieurs automatiques

3.1 Principes et choix généraux pour l'approche statistique

Afin d'exploiter au mieux la puissance de notre algorithme d'apprentissage automatique et d'obtenir les meilleurs résultats, nous avons eu recours à la validation croisée (80% vs 20%) pour tester plusieurs configurations. Pour des raisons de vitesse d'entrainement, seuls les 20000 premiers tweets ont été utilisés pour nos tests. L'algorithme d'apprentissage automatique supervisé retenu est *liblinear*[6] (Fan et al., 2008) avec les paramètres par défaut. Nous avons bien évidemment veillé à respecter la proportion de chaque catégorie pour la création de nos sous-corpus.

Afin d'obtenir un résultat optimal, nous avons testé plusieurs configurations grâce à la combinaison de différents traits. Nous avons ainsi utilisé : les unigrammes (lemmes, *tokens*), les *PoS*, les dépendances syntaxiques, les groupes nominaux d'une longueur maximale de 4, comme par exemple : "arrêt de bus", "station de métro", etc. et enfin les traits sémantiques qui reposent sur l'enrichissement sémantique.

Nos expérimentations ont montré que nous obtenons les meilleurs résultats en combinant ces différents traits : dépendances syntaxiques, enrichissement sémantique et groupes nominaux. Au regard des résultats, nous avons préféré écarter l'utilisation des *PoS* qui les faisaient décroitre. Nous expliquons ce phénomène en partie par le langage utilisé sur Twitter, qui est non standard, ce qui a tendance à fausser les analyses morphosyntaxiques des outils que nous utilisons.

3.2 Méthodes pour la tâche 1

Pour rappel, la tâche 1 avait pour objectif de distinguer les tweets évoquant les transports en commun des autres tweets. Nous avons mis en œuvre deux méthodes : une symbolique et une statistique. La symbolique se concentre sur l'enrichissement sémantique réalisé (en résumé, la présence de lexique spécifique à la thématique). En comparant les deux méthodes, nous cherchons à savoir si, dans le cas de tweets qui présentent par nature peu de variations, la puissance d'un algorithme d'apprentissage automatique qui prend en compte plusieurs traits, présente un réel intérêt.

3.2.1 Approche symbolique (run 1) : lean and mean

Notre piste de départ était de réaliser une méthode symbolique se basant sur le lexique spécifique au domaine des transports en commun. Nous avons rapidement réalisé que la manifestation de cette thématique dans les tweets est portée par le lexique et que la syntaxe notamment n'apporte pas d'information supplémentaire pour sa détection. Par exemple, dans la phrase "Je suis dans le bus", le *token* "bus" à lui seul permet de déterminer qu'il s'agit d'un tweet évoquant les transports. C'est généralement le cas pour la détection de classes thématiques. Cela est bien différent dans le cas de la tâche 2 comme nous l'avons déjà évoqué.

[6] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

Notre méthode symbolique est donc "*lean and mean*" (ou "bête et méchant") puisqu'elle consiste tout simplement à classer dans la catégorie *transport* tous les tweets comportant un *token* annoté TRANSPORT (cf. 2.2).

3.2.2 Approche statistique (run 2)

Nous avons réalisé un *run* en utilisant les paramètres préalablement définis (cf. 3.1), en particulier l'enrichissement sémantique TRANSPORT décrits en section 2.2. Cette approche nous a paru envisageable car nous avions un corpus annoté conséquent, ce qui représente un socle solide pour un algorithme d'apprentissage automatique.

3.3 Modèles pour la tâche 2

La tâche 2 proposait de détecter la polarité globale d'un tweet évoquant les transports en commun et de le classer parmi les catégories suivantes :

- *positif* : tweet à polarité uniquement positive (21% dans le corpus d'entrainement).
- *négatif* : tweet à polarité uniquement négative (37%).
- *neutre* : tweet sans opinions (36%).
- *mixposneg* : tweet contenant de la polarité positive et négative (7%).

Cette tâche est plus complexe que la première. En effet, il nous était difficile d'élaborer un système symbolique *lean and mean* en nous basant uniquement sur le lexique du corpus comme nous l'avons fait pour la tâche 1. D'autre part, comme nous l'avons expliqué en introduction, les méthodes symboliques d'analyse des sentiments dont nous disposons déjà, ont été développées pour le domaine de la relation client et ne donnent pas de bons résultats sur les tweets.

Nous avons alors décidé d'utiliser uniquement des méthodes statistiques pour cette tâche. Néanmoins, une des limites de cette méthode est la qualité du corpus utilisé pour l'entrainement des modèles. Dans notre cas, le corpus était quelque peu bruité. De même, certaines annotations étaient discutables. Les performances de l'algorithme d'apprentissage automatique ont pu donc être impactées de façon négative par ces différents facteurs. Nous en parlerons plus précisément dans la suite de l'article.
Nous avons alors réalisé une série de tests pour comparer différentes configurations. Nous avons opté pour la construction de deux modèles : pour chacun nous avons le même paramétrage, mais le corpus utilisé lors de l'apprentissage diffère.

3.3.1 Modèle avec la catégorie mixposneg

Le premier modèle a été construit avec l'intégralité du corpus annoté fourni, avec le paramétrage que nous avons décrit plus haut. Pour rappel, nous avons utilisé pour la création de ce modèle l'enrichissement sémantique grâce au lexique polarisé que nous avons extrait du corpus, ainsi qu'un lexique d'émojis polarisés (Novak et al., 2015). Ainsi, nous avons donné une polarité à la quasi majorité des émojis existants.

Nous avons remarqué, à la vue de nos résultats, que les tweets annotés en *mixposneg* étaient parfois très discutables. Il s'agit de plus de la catégorie la moins représentée (7%) et donc de celle pour laquelle nous avons le moins de données pour l'apprentissage du modèle. Lorsque nous entraînions notre modèle avec cette catégorie, nos résultats de précision, rappel et F-mesure tendaient à chuter.

Catégorie de référence	Nb d'occurrences	TP[7]	FP[8]	Précision
mixposneg	461	152	963	14%

TABLE 1 : précision de la catégorie *mixposneg* lors de la validation croisée

C'est pourquoi nous avons décidé d'entrainer un modèle sur un corpus sans la catégorie *mixposneg*. Notre hypothèse était que cette catégorie introduit trop de confusion et qu'il valait donc mieux ne pas l'utiliser.

3.3.2 Modèle sans la catégorie mixposneg

La validation croisée appliquée à un corpus d'entrainement privé de la catégorie *mixposneg* nous a permis d'obtenir de meilleurs résultats (F-mesure, Rappel, Précision). Ces résultats sont logiques, puisque la précision du repérage de la catégorie *mixposneg* est extrêmement basse (cf. table 1) et a un impact négatif sur nos résultats totaux.

Nous avons testé nos deux modèles (avec ou sans *mixposneg* dans le corpus d'apprentissage) sur une partie du corpus d'entrainement réservé à être utilisé comme corpus de test. Ce dernier était alors composé de tous les labels de référence, comprenant également la catégorie *mixposneg*. Nous avons alors obtenu les résultats ci-contre[9].

Configuration des modèles	Précision	Rappel	F-Mesure
Corpus d'entrainement sans *mixposneg*	0,656	0,656	0,656
Corpus d'entrainement avec *mixposneg*	0,643	0,643	0,643

TABLE 2 : comparatif de résultats entre les deux modèles appliqués au même fichier de test (avec *mixposneg*)

Ces résultats se confirment lors de l'évaluation finale (cf Table 3), notre modèle ne comprenant pas de *mixposneg* (*run* 1) étant plus performant que le modèle entrainé avec (*run* 2).

4 Evaluation

Cette section décrit les performances de toutes les équipes aux tâches 1 et 2. Onze équipes ont soumis un total de 38 *runs* à la tâche 1, et 39 *runs* à la tâche 2.

7 TP : *true positive* (vrai positif)

8 FP : *false positive* (faux positif)

9 Les résultats obtenus sont issus de notre propre formule qui diffère de celle utilisée pour le DEFT.

4.1 Evaluation pour la tâche 1

Le tableau ci-dessous récapitule les performances des meilleurs *runs* pour chaque équipe à la tâche 1. Pour notre équipe, nous donnons les chiffres pour nos deux *runs*. Les valeurs marquées avec * ont une différence statistiquement significative[10] avec notre meilleur *run*, avec une valeur-p de 0,01. Les valeurs marquées avec ** sont statistiquement différentes avec une valeur-p de 0,05.

Rang[11]	**N° équipe_n° tâche**	***Run***	**Précision**	**Rappel**	**F1**
1	3_T1	2	0,83124	1	0,90785*
2	7_T1	5	0,83048	1	0,90739*
6	6_T1	4	0,82702	1	0,90532**
11	5_T1	1	0,82433	1	0,90371
12	1_T1	1	0,82293	1	0,90286
18	**8_ELOQUANT_T1**	**1**	**0,81408**	**0,99984**	**0,89745**
19	11_T1	3	0,80604	1	0,8926
21	9_T1	3	0,80502	1	0,89198
MOYENNE			*0,80293*	*0,99997*	*0,89029*
29	10_T1	3	0,79580	1	0,88629*
30	**8_ELOQUANT_T1**	**2**	**0,79360**	**0,99984**	**0,88486***
33	2_T1	2	0,77955	1	0,87612*
37	14_T1	1	0,70509	1	0,82704*

TABLE 3 : performances des équipes à la tâche 1 du DEFT 2018.

La moyenne des F-mesures des 38 *runs* est 0,8882. La plupart des équipes obtient des scores très comparables dans cette tâche (c'est moins le cas pour la tâche 2).

Notre meilleur *run*, correspondant à l'approche symbolique, est au-dessus de la moyenne, avec un F-score de 0,89745, et l'écart entre cette performance et le champion est faible (0,014 point de F-mesure). L'écart entre nos deux *runs* est statistiquement significatif, et d'environ 0,013 points de F-mesure.

4.2 Evaluation pour la tâche 2

Le tableau ci-dessous récapitule les performances des meilleurs *runs* à la tâche 2. Pour notre équipe, nous donnons les chiffres pour nos deux *runs*. Nous avons omis le *run* d'une équipe qui présentait une anomalie (valeurs numériques `NaN`). Les résultats des tests de significativité sont représentés de la même façon que pour la tâche 1.

[10] Selon le test de Ficher

[11] Parmi tous les *runs*

Rang[12]	N° equipe_n° tâche	*Run*	Micro-précision	Micro-rappel	Micro-F1
1	5_T2	5	0,69906	1	0,82288*
4	6_T2	1	0,70258	0,96497	0,81313**
6	3_T2	2	0,67699	1	0,80738
9	1_T2	3	0,67013	1	0,80249
11	**8_ELOQUANT_T2**	**1**	**0,66684**	**1**	**0,80012**
13	**8_ELOQUANT_T2**	**2**	**0,66151**	**1**	**0,79627**
15	9_T2	4	0,64172	1	0,78176*
17	10_T2	1	0,63004	1	0,77304*
19	7_T2	5	0,65824	0,93075	0,77113*
MOYENNE			*0,59364*	*0,95199*	*0,72538**
31	15_T2	3	0,47628	1	0,64524*
37	2_T2	1	0,34255	1	0,5103*
39	14_T2	1	0,29778	0,52821	0,38085*

TABLE 4 : performances des équipes à la tâche 2 de DEFT 2018

La moyenne des F-mesures des 39 *runs* est d'environ 0,73, trois quarts des *runs* ayant une F-mesure supérieure à 0,7. Notre meilleur *run*, correspondant au classifieur entraîné en l'absence de tweets annotés comme *mixposneg*, produit une F-mesure de 0,80. Cela place notre système dans les meilleurs 30% de tous les *runs* pour la tâche 2. L'écart entre ce *run* et le *run* champion est d'environ 0,023 points de F1, et c'est le seul système avec lequel nous observons une différence statistiquement significative (valeur-p à 0.01). L'écart entre nos deux *runs* est d'environ 0,004 points, cette différence n'étant pas significative.

5 Discussion

5.1 Choix et intérêts des méthodes

Pour la tâche 1, nous observons que seulement les deux premiers systèmes ont une performance statistiquement supérieure à celle de notre meilleur *run* (valeur-p = 0.01). En fait, pour cette tâche, la plupart des systèmes ont des résultats comparables et assez élevés. Nous pensons que cela est dû au fait que la tâche est relativement simple, et que, comme nous l'avons vu avec le système symbolique, la distinction peut être fortement basée sur le lexique.

L'approche symbolique a une performance légèrement meilleure que celle de l'approche par apprentissage automatique. Au cours des évaluations que nous avons menées lors des développements, nous avons observé quelques différences intéressantes entre ces deux méthodes. Nos résultats par validation croisée sur le corpus d'entrainement ont notamment présenté un rappel pour la catégorie *transport* plus élevé avec la méthode symbolique. Celle-ci étant basée sur une sélection du lexique

[12] Parmi tous les *runs*

spécifique aux transports en commun, elle permet de bien détecter les tweets de la thématique et de ne pas en laisser de côté. La nature de la tâche 1, qui est nous semble-t-il basique, se prête bien à ce type de méthode *lean and mean* telle que décrite en 3.2.1.

A l'issue de la période de test, nous avons également cherché à comprendre la différence entre nos deux méthodes pour la tâche 1 en observant les tweets pour lesquels les deux *runs* produisaient des sorties différentes. Les cas où la méthode statistique attribue à tort la catégorie *transport*, concernent des tweets qui abordent des sujets proches des transports en commun mais sans les évoquer directement (par exemple : travaux, circulation, horaires, voyages). Ce phénomène est difficile à corriger. A l'inverse, lorsque l'approche statistique attribue à raison la catégorie *transport*, et non l'approche symbolique, cela est dû à sa capacité à prendre en compte des traits variés. Par exemple, lorsqu'une entité nommée était mentionnée par une forme non identifiée par notre lexique : pour le RER E, nous avions listé "RER", "RER_E", "RERE_RATP" mais pas "RERE_SNCF". Un travail de normalisation pourrait corriger cela dans la méthode symbolique. Le fait d'observer les deux méthodes, permet de trouver des pistes pour améliorer les deux. Une poursuite intéressante de ce travail serait de développer une méthode hybride combinant l'intérêt de chacune d'elles, et corrigeant mutuellement leurs faiblesses.

Pour la tâche 2, nos deux *runs* sont très proches et, en prenant en compte les tests de significativité, très bien situés dans le classement global. Même si la différence entre nos deux *runs* n'est pas statistiquement significative, il est intéressant de noter que les meilleurs résultats sont obtenus en ignorant l'annotation *mixposneg* pour entrainer le modèle. Dans ce *run*, le modèle ne va jamais assigner cette catégorie à un tweet, donc ce système a toujours une Précision et un Rappel égale à 0 pour cette catégorie. Néanmoins, les résultats montrent qu'il est préférable d'accepter ces erreurs systématiques que d'essayer de créer un modèle capable de reconnaitre cette catégorie. En effet, la très basse fréquence de cette catégorie dans le corpus et la grande ambiguïté qu'elle présente introduisent trop de bruit pour les systèmes d'apprentissage automatique.

5.2 Qualité du corpus

Les textes issus de médias sociaux, et particulièrement de ceux qui imposent des contraintes sur la longueur des énoncés, présentent des phénomènes morpho-lexicaux et morphosyntaxiques peu standard. De plus, Twitter, étant une plate-forme privilégiée par une population particulière, a développé un sociolecte particulier, comprenant des lexiques, des graphies, et des tournures de phrase propres à ce milieu. La conséquence de cela est l'inadéquation d'outils d'analyse du français plus standard pour le traitement de ces textes.

Le corpus est d'une taille très conséquente, ce qui pousse à penser que, une fois des normalisations effectuées, la quantité des données est suffisante pour que des méthodes classiques d'apprentissage automatique s'approchent de leurs performances maximales théoriques.

Cependant, en parcourant les annotations nous avons constaté des imperfections. Pour de nombreux tweets notre équipe est unanime sur le fait que l'annotation ne correspond pas à la valence émotionnelle ressentie. Le tableau ci-dessous donne quelques exemples ; certains cas de figure sont plus fréquents que d'autres.

Label de référence	Label « ressenti »	Tweet
positif	*négatif*	@nighshxde c' est à cause de ma 4g aussi :(si elle marchait dans le métro Ca m' arrangerai 😭😭😭😭😭😭😭
positif	*négatif*	Avec la putain de rentrée des classes , mon bus est arrivé en retard donc j' ai loupée mon train. Super. Génial.
mixposneg	*négatif*	Panne du ter17758 arrêté au milieu des voies. On cuit dedans avec cette chaleur et tout est fermé déjà 1h de retard merci #SNCF @SNCF
mixposneg	*négatif*	@LIGNEL_sncf pas de bus 275 et pas de train , et évidemment pas de remboursement merci @SNCF fdp
négatif	*neutre*	@SNCF Hello , le site de la SNCF étant en maintenance , pouvez-vous m' indiquer sur le train Lyon/Paris n°6616 de dimanche est annulé ? Merci.
neutre	*négatif*	Question aux parents : il fait 300° dans le bus et 2 gamins se sont mis à chanter 🎶Libérée , délivrée🎶. Ai-je le droit de les tuer ? 🤔 Merci !
neutre	*négatif*	Eh le conducteur du bus ! Le bus c' est pas un kart alors calmes toi !
positif	*neutre*	Retour à un trafic régulier sur l' ensemble de la #Ligne8 #RATP. Incident terminé
négatif	*positif ?*	A Paname je suis choquée ya des prises sur les abris de bus 😐

TABLE 5 : Exemples d'annotations polémiques

Le plus gros problème constaté est l'interprétation quasi-littérale des tweets, l'ironie n'étant que rarement remarquée par les annotateurs. Or elle semble très présente dans les tweets fournis. Toutefois, il ne s'agit pas ici de minimiser l'effort et la difficulté de la tâche d'annotation, car pour certains tweets nous-mêmes n'étions d'accord ni entre nous, ni avec le label de référence.

Il y a quelques rares cas de tweets dupliqués et annotés différemment, comme le suivant, annoté comme *neutre* et *négatif* :

```
"@RERE_SNCF 3€ du + sur le navigo , ça permettra d' avancer autrement
qu' au ralenti avant Villiers pour ne pas rater le bus ? #qml #marre".
```

Il est difficile de quantifier l'impact de ce bruit sur les scores, mais il est clair qu'ils ne peuvent qu'en être dégradés.

5.3 Améliorations possibles

Une étape essentielle du prétraitement des textes bruités est la normalisation. Il existe différentes approches, dont un état de l'art peut être trouvé dans (Tarrade, 2017). Ces approches vont de simples substitutions lexicales prédéfinies, à la traduction automatique (Kaufmann et Kalita, 2010).

Actuellement, nous effectuons une normalisation minimale, qui peut être améliorée. Han et Baldwin (2011) proposent de réduire à deux lettres toute suite de plus de deux lettres identiques à l'intérieur d'une forme (ainsi 'coooool' devient 'cool'). Une approche similaire, dans laquelle on réduirait à une lettre permettrait de ramener des formes 'ptdrrrr', 'looool', 'nooooooon' aux plus fréquentes 'ptdr', 'lol', 'non', respectivement. Nous effectuons déjà un traitement *ad hoc* de certains de ces cas, mais n'avons pas encore généralisé cette approche.

Nous pourrions également exploiter les ressources lexicales de normalisation de langage SMS et les outils de normalisation développés sur des corpus proches (Tarrade et Lopez, 2017), comme le corpus 88milSMS (Panckhurst et al., 2014).

Alternativement, une façon non supervisée d'approcher les données serait d'utiliser des vecteurs de mots word2vec (Mikolov et al, 2013) construits sur un grand ensemble de tweets en français.

Enfin, étant donné les désaccords constatés autour des annotations, une ré-annotation collaborative du corpus pourrait être envisagée. Cela permettrait de marquer en même temps les tweets contenant de l'ironie, caractérisant ainsi quantitativement ce phénomène linguistique difficile à identifier.

6 Conclusion

Nous avons présenté les systèmes de l'équipe Eloquant pour la catégorisation de tweets en français dans les tâches 1 (détection de la thématique transports en commun) et 2 (détection de la polarité globale) du DEFT 2018.

En adaptant nos outils existants, reposant sur un enrichissement sémantique, l'apprentissage automatique et, pour la tâche 1, une approche symbolique, nous avons obtenu des F-mesures au-dessus de la moyenne globale pour chaque tâche, nous plaçant dans les meilleurs 30% de tous les *runs* pour la tâche 2. Il est également notable que pour cette tâche, seul le système avec la meilleure performance a une différence statistiquement significative avec celle de notre système. Cela implique que notre système est en réalité comparable aux systèmes classés en deuxième position, ce que nous considérons un très bon résultat.

Nous considérons très satisfaisants les résultats obtenus dans les deux tâches. Les systèmes que nous avons présentés ont été adaptés à partir du système que nous utilisons habituellement pour le domaine de la relation client (Maurel et al., 2008), qui traite un type de langage assez différent de celui de Twitter et des opinions sur les transports en commun. Ainsi, être comparable aux systèmes qui obtiennent les deux ou trois meilleurs résultats nous parait une indication de la maturité de nos systèmes d'analyse sémantique, qui ont pu être adaptés à ce nouveau domaine avec une intervention assez minimale.

Références

ABDAOUI, A., TAPI NZALI, M. D., AZE, 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. *Présenté à 22ème Traitement Automatique des Langues Naturelles*, Caen, France.

CHEN, M., JIN, X., & SHEN, D. (2011). Short Text Classification Improved by Learning Multi-granularity Topics. *In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three* (p. 1776–1781). Barcelona, Catalonia, Spain: AAAI Press.

FAN, R.-E., CHANG K.-W., HSIEH C.-J., WANG X.-R., ET LIN C.-J. (2008). LIBLINEAR : A library for large linear classification. *Journal of machine learning research n° 9*, Aug. 1871-1874.

GUIBON, G., OCHS, M., & BELLOT, P. (2016). From Emojis to Sentiment Analysis. *In WACAI* 2016.

HAN, B., BALDWIN, T. (2011). Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter. *In Proc eedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (p. 368 – 378). Stroudsburg, PA, USA: Association for Computational Linguistics.

KAUFMANN, M., KALITA, J. (2010). Syntactic normalization of twitter messages. *In International conference on natural language processing,* Kharagpur, India.

MAUREL, S., CURTONI, P., & DINI, L. (2008). A hybrid method for sentiment analysis. *In INFORSID*. http://www.ho2s.com/assets/celi-france_english-2.pdf

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., DEAN, J. (2013). Distributed representations of words and phrases and their compositionality. *In Advances in neural information processing systems* (pp. 3111-3119).

NOVAK, P. K., SMAILOVIĆ, J., SLUBAN, B., & MOZETIČ, I. (2015). Sentiment of emojis. *PloS one, 10(12), e0144296.*

PANCKHURST, R., DETRIE, C., LOPEZ, C., MOÏSE, C., ROCHE, M., VERINE, B. (2014). Un grand corpus de SMS en français : 88milSMS.

PAROUBEK, P., GROUIN, C., BELLOT, P., VINCENT CLAVEAU, ESHKOL-TARAVELLA, I., FRAISSE, A., JACKIEWICZ, A., KAROUI, J., MONCEAUX, L., TORRES-MORENO, J.-M. (2018). DEFT2018 : recherche d'information et analyse de sentiments dans des tweets concernant les transports en Île de France. *In Actes de DEFT*. Rennes, France.

SCLANO, F., & VELARDI, P. (2007). TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. *In Enterprise Interoperability II* (p. 287-290). Springer London.

TARRADE, L., LOPEZ, C. (2017). Corpus de tweets et de SMS annotés pour l'observation de phénomènes linguistiques en français "non standard". *Actes TALN'2017.*

TARRADE, L. (2017). Normalisation des messages issus de la communication électronique médiée. *Mémoire de M2R. Sciences de l'Homme et Société.* 〈dumas-01666146〉

URIELI, A. (2013). Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit. *Thèse de doctorat. Université Toulouse le Mirail-Toulouse II.*

VERNIER, M., MONCEAUX, L., DAILLE, B., & DUBREIL, E. (2009). Catégorisation des évaluations dans un corpus de blogs multi-domaine. *Revue des Nouvelles Technologies de l'Information*, 45--70.