Système d'ensemble pour la classification de tweets, DEFT2017

Damien Sileo[1, 2] Camille Pradel[1,*] Philippe Muller[2,*] Tim Van de Cruys[2,*]
(1) Synapse Développement, 5 Rue du Moulin Bayard, 31000 Toulouse
(2) IRIT, Université Paul Sabatier 118 Route de Narbonne 31062 Toulouse
(*) Contributions égales
damien.sileo@synapse-fr.com, camille.pradel@synapse-fr.com,
philippe.muller@irit.fr, tim.van-de-cruys@irit.fr

RÉSUMÉ

Nous présentons le système utilisé par l'équipe MELODI dans la compétition DEFT2017 portant sur la détection de langage figuratif et de polarité de tweets en français. En se basant sur de l'apprentissage non supervisé, un système unique et générique est proposé qui surpasse la moyenne des systèmes sur les trois tâches en n'utilisant que des informations lexicales et sans ressource structurée extérieure.

ABSTRACT

Ensemble system for tweets classification, DEFT2017

We present the system used by the MELODI team in the DEFT2017 =competition which addresses detection of figurative language and polarity of tweets in French. We propose a single and generic system based on unsupervised learning, which scored above the average of the systems on the three tasks, using only lexical information and no structured external resource.

MOTS-CLÉS : *Fasttext*, langage figuratif, analyse de sentiment.

KEYWORDS: *Fasttext*, figurative language, sentiment analysis.

1 Introduction

L'analyse de sentiments permet de dégager automatiquement des indicateurs spécifiques sur une grande quantité de textes. Un certain nombre de phénomènes linguistiques compliquent la tâche. Parmi ceux-ci, le langage figuratif détourne le sens des phrases d'une interprétation naïve. La compétition DEFT2017 (Benamara *et al.*, 2017) mêle les aspects d'analyse de sentiments et de détection de langage figuratif en proposant trois tâches liées à des tweets en français :

1. La classification de tweets non figuratifs selon leur polarité ;
2. L'identification de langage figuratif selon les catégories d'ironie, sarcasme, humour ou non-figuratif ;
3. La classification des tweets figuratifs et non figuratifs selon leur polarité.

On propose d'évaluer l'apport de représentations lexicales dans ces tâches, avec un seul système avec la même configuration d'hyperparamètres sur les trois. Deux types de représentations sont combinées : des représentations parcimonieuses (bigrammes tf-idf) et continues (représentations *Fasttext*).

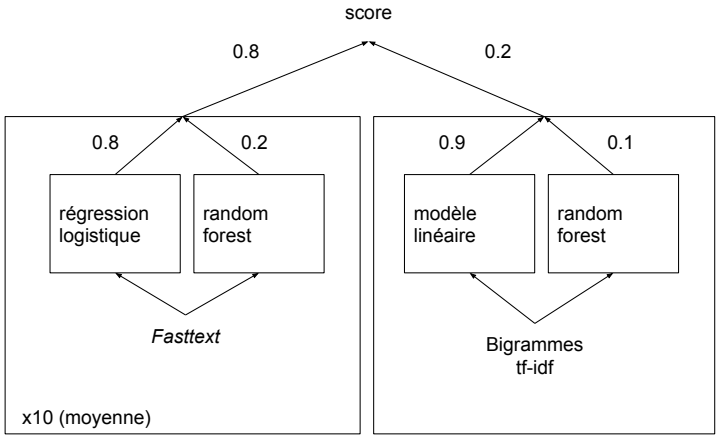

FIGURE 1 – architecture du système proposé

2 Modèle

2.1 Vue d'ensemble

Deux ensembles d'indicateurs sont utilisés :

— Des représentations tf-idf de bigrammes ;
— Des moyennes de représentations vectorielles des mots de la phrase obtenues avec *Fasttext* (Bojanowski *et al.*, 2016).

Chacun de ces ensembles est donné en entrée à deux classifieurs, et la probabilité attribuée à un score est une combinaison linéaire des probabilités fournis par les classifieurs. La section explique l'obtention de coefficients.

Les classifieurs utilisant les indicateurs de *Fasttext* sont entraînés 10 fois avec des modèles *Fasttext* identiques mais ré-appris à chaque fois (*Fasttext* étant stochastique).

La figure 1 montre l'architecture utilisée. Les flèches accompagnées d'un nombre correspondent aux sommes pondérées de probabilités.

2.2 *Fasttext*

Le modèle skipgram *Fasttext*(Bojanowski *et al.*, 2016) est basé sur le modèle skipgram de *word2vec* (Mikolov *et al.*, 2013), qui consiste à apprendre des représentations de mots pour qu'elles optimisent une tâche de prédiction du contexte des mots, illustrée à la figure 2. La différence principale est que la représentation h_w d'un mot w ne se résume plus à u_w, la représentation de son symbole. Elle est

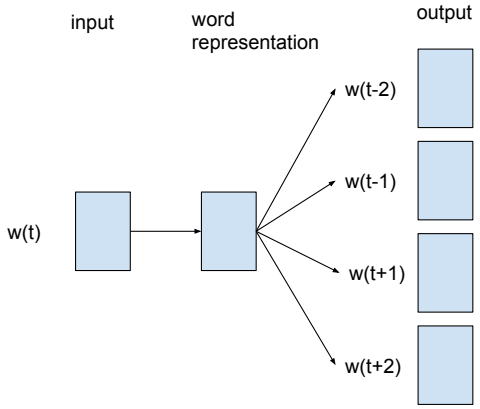

FIGURE 2 – tâche d'apprentissage du modèle skipgram

augmentée de la représentation des n-grammes de caractères contenus dans w, nommés $u_g, g \in \mathcal{G}_w$:

$$h_w = u_w + \sum_{g \in \mathcal{G}_w} u_g \tag{1}$$

$\mathcal{G}_w$ correspond aux n-grammes de w suffisamment fréquents et d'une taille adéquate. La morphologie de w est donc partiellement prise en compte dans h_w, même si l'ordre des grammes est ignoré.

3 Expériences

Pour apprendre les représentations de mots de *Fasttext*, nous avons utilisé des tweets issus de la plateforme *OSIRIM* [1] de l'IRIT qui collecte 1% du flux de Twitter depuis Septembre 2015. Nous avons sélectionné un échantillon de ces tweets et ignoré 99% des tweets contenant un hyperlien, sachant que bon nombre d'entre eux proviennent de bots. 1% des mentions (*@someone*) sont remplacés par le symbole @* en espérant améliorer ainsi la généralisation. L'ensemble résultant totalise $40M$ tweets. Les paramètres utilisés par *Fasttext* sont résumés dans la table 1.

3.1 Méthodologie et paramètres

La sélection de modèle s'est effectuée par validation croisée sur la tâche 3, en maximisant la métrique F1-micro. Les modèles et paramètres sont sélectionnés de manière gloutonne : nous avons optimisé séparément les deux modèles retenus par ensemble d'indicateurs, puis les coefficients de combinaison des modèles utilisant le même ensemble d'indicateurs, et enfin la combinaison finale.

Le modèle linéaire basé sur les bigrammes utilise une fonction de coût *hinge* et une pénalité Elasticnet (Zou & Hastie, 2005) de coefficient 10^{-4} et dont le ratio l_1 est de 0.05. Le modèle linéaire est calibré

1. http://osirim.irit.fr/site/fr/articles/corpus

paramètres	**valeur**
learning rate	0.02
dimensions	200
context window size	5
epochs	4
min_count	5
negative/positive samples ratio	5
loss	negative sampling
minimum n-gram size	3
maximum n-gram size	6
sampling threshold	10^{-4}

TABLE 1 – Paramètres du modèle *Fasttext*

tâche	**ensemble**	**moyenne**	**médiane**	**min**	**max**
1	0.546	0.475	0.523	0.239	0.645
2	0.702	0.694	0.720	0.476	0.783
3	0.517	0.473	0.519	0.231	0.594

TABLE 2 – Mesures F1-macro sur les différentes tâches, comparées aux autres compétiteurs

avec un modèle isotonique pour produire des estimations de probabilité. Les modèles random forest (Breiman, 2001) utilisent 200 estimateurs avec les indicateurs bigrammes et 300 estimateurs avec les indicateurs *Fasttext*.

4 Evaluation

Le tableau 2 présente les résultats fournis par le système d'évaluation. Notre système est désigné par *ensemble*, *moyenne* désigne la moyenne des résultats parmi les participants, et ainsi de suite. Les résultat sont proches de la médiane et au dessus de la moyenne à chaque fois ce qui démontre une certaine robustesse. L'évaluation du système utilise la métrique F1-macro alors que la métrique F1-micro a été utilisée pour l'optimisation ce qui est certainement dommageable pour les résultats.

5 Conclusion

Nous avons décrit un système présenté à la compétition DEFT2017. L'approche purement lexicale donne de bons résultats mais reste limitée face à la complexité inhérente des tâches, qui mélangent divers aspects sémantiques et contextuels. Prendre en compte les spécificités de la tâche pourrait sans doute améliorer les résultats, mais en restant dans des modèles génériques, il serait intéressant d'évaluer l'apport de méthodes d'apprentissage non supervisées traitant l'ordre des mots, telles que *Skipthought* (Kiros *et al.*, 2015).

Références

BENAMARA F., GROUIN C., KAROUI J., MORICEAU V. & ROBBA. I. (2017). Analyse d'opinion et langage figuratif dans des tweets : présentation et résultats du Défi Fouille de Textes DEFT2017. *Actes de l'atelier DEFT de la conférence TALN 2017, 26 juin 2017, Orléans.*

BOJANOWSKI P., GRAVE E., JOULIN A., MIKOLOV T., GRAVE E., BOJANOWSKI P., MIKOLOV T., LAKE B. M., ULLMAN T. D., TENENBAUM J. B., GERSHMAN S. J. & SCHMIDHUBER J. (2016). Bag of Tricks for Efficient Text Classification. *arXiv :1604.00289v1[cs.AI]*, p. 1–55.

BREIMAN L. (2001). Random forest. *Machine Learning*, **45**(5), 1–35.

KIROS R., ZHU Y., SALAKHUTDINOV R., ZEMEL R. S., TORRALBA A., URTASUN R. & FIDLER S. (2015). Skip-Thought Vectors. *ArxiV*, (786), 1–11.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Nips*, p. 1–9.

ZOU H. & HASTIE T. (2005). Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society*, **67**(2), 301–320.