Cameli@ : analyses automatiques d'e-mails pour améliorer la relation client

Guillaume Dubuisson Duplessis Sofiane Kerroua Ludivine Kuznik
Anne-Laure Guénet
EDF Commerce, Direction Numérique, Tour PB6, 178 Rond-Point de la Défense, 92800 Puteaux, France
anne-laure.guenet@edf.fr

RÉSUMÉ

Cette démonstration présente un système actuellement en production d'analyses automatiques d'e-mails en français incluant des analyses thématiques, des analyses de l'opinion, des tâches d'extraction d'information et une tâche de pseudo-anonymisation.

ABSTRACT

Cameli@ : automatic e-mail analysis to improve the customer relationship

This demonstration presents an automatic e-mail analysis system in French language including thematic analysis, opinion analysis, information extraction tasks and a pseudo-anonymization task.

MOTS-CLÉS : classification thématique, analyse de l'opinion, extraction d'information.

KEYWORDS: thematic classification, opinion mining, information extraction.

1 Analyses automatiques des e-mails client à EDF Commerce

Chaque mois plusieurs centaines de milliers d'e-mails client sont envoyés à EDF et font l'objet d'un traitement manuel attentif de la part des conseillers client. Ces données textuelles en français offrent une grande richesse. Résultat d'une expression libre et spontanée du client, elles font notamment apparaître une grande variabilité, par exemple, en terme de nombre de sujets abordés, de respect de l'orthographe et de la syntaxe, de niveau de langue, de politesse mais aussi de structuration. L'augmentation significative de l'utilisation des canaux digitaux comme les e-mails appelle à une optimisation du processus de réponse afin de raccourcir le temps de traitement des e-mails tout en assurant un très haut niveau de qualité dans la relation client. C'est pour faire face à ce défi que le projet Cameli@ a vu le jour. Il met en place des technologies récentes du domaine du TALN en production afin d'optimiser le traitement des e-mails clients pour l'amélioration de la relation client.

Les analyses automatiques de Cameli@ visent à enrichir « à la volée » les e-mails d'information comme la thématique de contact ou le ressenti du client afin de nourrir trois cas d'usage. Premièrement, elles ont été pensées afin d'optimiser le routage vers le conseiller le plus adapté. Ensuite, les résultats des analyses sont exposés via un tableau de bord permettant le pilotage de l'activité de traitement des e-mails. Nourrie au fil de l'eau, cette interface permet aux directions métier de suivre les pics de contacts sur les différents motifs et ainsi gagner en réactivité dans leurs process. Aussi, il permet d'avoir une vision allant de plusieurs mois à la journée courante tout en garantissant un retour aux e-mails pseudo-anonymisés. Enfin, les analyses permettent l'automatisation du traitement de

certaines demandes simples et ciblées afin de décharger les conseillers client. Par exemple, un e-mail client informant uniquement d'une relève d'index peut faire l'objet d'un traitement complètement automatique.

Le projet Cameli@ met en place des analyses variées dans un cadre opérationnel. Il a été pensé dans le respect de contraintes de nature diverse. La première contrainte est de nature légale. Les analyses et les cas d'usage ont été conçus en prenant en compte nativement la loi RGPD. Par exemple, un algorithme de pseudo-anonymisation des e-mails a été conçu pour permettre un retour aux données dans le tableau de bord occultant des éléments identifiants tels que les prénoms, noms et adresses postales. La deuxième contrainte est de nature technique. Les analyses sont développées pour un fonctionnement en production dans l'écosystème de la DSI Commerce d'EDF. Cela impacte notamment le choix des langages de programmation ou encore l'infrastructure cible (par exemple, la présence ou non de GPU). En outre, le temps de traitement des analyses est optimisé pour permettre un enrichissement des e-mails à la volée au fur et à mesure de leur réception. Enfin, la troisième contrainte est une exigence « métier » en terme de performances et d'un certain niveau de transparence des modèles (Weller, 2017). Effectivement, les résultats des modèles sont exploités par des utilisateurs qui ne sont pas les concepteurs des modèles. Afin de garantir leur utilisabilité, il est nécessaire d'assurer un niveau de performance optimal et leur capacité à généraliser. Il est aussi important de donner une certaine intuition du fonctionnement du modèle, en particulier pour les modèles de type « boîte noire ». Dans ce contexte, le retour aux données est primordial.

Le projet Cameli@ met en oeuvre trois grands types de tâches. Il inclut de multiples tâches de classification. Tout d'abord, une quinzaine de catégorisations thématiques non-exclusives qui permettent de cerner le contenu de l'e-mail (par exemple, des catégories telles que « réclamation », « coupure », « montant de la facture »). Ensuite, des modèles d'analyse de l'opinion (Clavel *et al.*, 2013) afin de cerner la polarité émotionnelle de l'e-mail telle que la présence de l'expression d'un mécontentement. Le projet inclut des tâches d'extraction d'information dans le but d'automatiser le traitement des demandes simples. À l'heure actuelle, il s'agit de l'extraction des index du compteur communiqués par e-mail. Enfin, le projet implique une tâche de pseudo-anonymisation qui vise à désidentifier les e-mails en supprimant des éléments identifiants tels que les noms et adresses postales. Cette dernière se fonde sur une tâche de reconnaissance d'entités nommées (Nouvel *et al.*, 2016).

Le projet a bénéficié de la mise en place d'un processus d'annotation rigoureux. Les données d'apprentissage ont été annotées en interne via une plateforme web inspirée du projet Camomile (Poignant *et al.*, 2016) avec un effort particulier pour assurer la qualité des annotations (triple annotation, accord inter-annotateur). L'importante quantité de données non-annotées a été exploitée par l'usage de modèles de plongement vectoriel entraînés spécifiquement sur les e-mails tels que Word2Vec (Mikolov *et al.*, 2013), FastText (Bojanowski *et al.*, 2016), et GloVe (Pennington *et al.*, 2014). De multiples approches de classification ont été explorées. Notamment, des approches neuronales de type CNN et LSTM. *In fine*, des classifieurs plus « simples » mais offrant des performances compétitives ont été retenus (Shen *et al.*, 2018). L'extraction d'information repose sur le logiciel XIP développé par la société Xerox fonctionnant à base de règles linguistiques.

2 Démonstrateur

Cette démonstration vise à exposer les différentes analyses réalisées sur les e-mails et à présenter le cas d'usage « tableau de bord ». À cet effet, le démonstrateur inclut (i) la possibilité pour l'utilisateur

de saisir un e-mail fictif, de lancer le traitement sur cet e-mail en « temps réel » et de visualiser la restitution des différentes analyses, et (ii) une présentation du tableau de bord qui permet de restituer et synthétiser les résultats des analyses sur une quantité importante d'e-mails.

Remerciements

Nous remercions chaleureusement toutes les personnes qui sont intervenues sur ce projet : Meryl Bothua, Sylvain Boucault, Christophe Malnou, Bulent Burgucuoglu, Mélanie Cazes, Ariane De Moegen, Thomas Desmettre, Uyen-To Doan-Rabier, Asceline Goudjo, Uta Hosokawa, Cécile Legrand, Youcef Maamra, Ibtissem Menacer, Mathilde Poulain, Jean-Charles Rue, Jérôme Simoneto et son équipe, Véronique Ubério, Jérémy Vialaneix, et Jean Vidal.

Références

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv :1607.04606*.

CLAVEL C., ADDA G., CAILLIAU F., GARNIER-RIZET M., CAVET A., CHAPUIS G., COURCINOUS S., DANESI C., DAQUO A.-L., DELDOSSI M. *et al.* (2013). Spontaneous speech and opinion detection : mining call-centre transcripts. *Language resources and evaluation*, **47**(4), 1089–1125.

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

NOUVEL D., EHRMANN M. & ROSSET S. (2016). *Named Entities for Computational Linguistics*. John Wiley & Sons.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.

POIGNANT J., BUDNIK M., BREDIN H., BARRAS C., STEFAS M., BRUNEAU P., ADDA G., BESACIER L., EKENEL H., FRANCOPOULO G. *et al.* (2016). The Camomile collaborative annotation platform for multi-modal, multi-lingual and multi-media documents. In *LREC 2016 Conference*.

SHEN D., WANG G., WANG W., MIN M. R., SU Q., ZHANG Y., LI C., HENAO R. & CARIN L. (2018). Baseline needs more love : On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, volume 1, p. 440–450.

WELLER A. (2017). Challenges for transparency. *arXiv preprint arXiv :1708.01870*.