Une approche stylométrique pour la fouille d'opinion

Gaël Lejeune, Frédéric Dumonceaux
(1) LINA, 2 rue de la Houssinière, 44322 Nantes, France
prenom.nom@univ-nantes.fr

Résumé. Dans cet article nous proposons une approche stylométrique pour l'édition 2015 du Défi Fouille de Textes. Cette édition du défi portait sur l'analyse d'opinions, de sentiments et d'émotions dans un corpus issu de *Twitter*. Nous avons participé dans trois tâches du défi : classification des *tweets* selon leur polarité (Tâche 1, 3 classes), identification de la classe générique de l'information exprimée dans le *tweet* (Tâche 2.1, 4 classes) et identification de la classe spécifique de l'opinion, sentiment ou émotion exprimée dans le *tweet* (Tâche 2.2, 18 classes). L'approche stylométrique que nous avons utilisée est fondée sur l'utilisation de n-grams de caractères de manière à traiter ces tâches de fouille d'opinion comme des tâches d'attribution d'auteur. Notre hypothèse était la suivante : les traits qui permettent de caractériser le style d'un auteur devraient permettre d'identifier le style inhérent à une classe d'opinion, de sentiment ou d'émotion. Finalement, cette hypothèse s'est avérée erronée, particulièrement sur la tâche 3 qui était la plus difficile. La première interprétation que l'on peut faire serait qu'il n'existe pas véritablement de traits stylistiques inhérents aux classes étudiées. Une autre explication possible est la faible longueur des messages qui rendrait les méthodes stylométriques inopérantes.

Abstract.

A stylometric approach for opinion mining

This article tries to tackle the DEFT'15 opinion mining challenge using a stylometric approach. The dataset proposed by the organizers was a set of microblog messages extracted from Twitter. We participated in three tasks : classification according to polarity (Task 1, 3 classes), classification according to information (Task 2.1, 4 classes) and classification according to specific classes (Task 3, 18 classes). The stylometric approach we used was based on recent work on Autorship Attribution using character n-grams as features. Our assumption was that the features efficient for characterizing an author style would be efficient as well for identifying the opinions or emotions expressed in tweets. We showed that this assumption was wrong, especially on task 3. It appears that the stylometric features might not be well suited for opinion mining tasks. Another hypothesis to explain this result is that the length of the microblog messages might be too small to take advantage of such a stylometric approach.

Mots-clés : stylométrie, attribution d'auteur, analyse d'opinion, analyse de sentiment, classification, chaînes de caractères, microblogs, tweets.

Keywords: stylometry, autorship attribution, opinion mining, sentiment analysis, classification, character substrings, microblogs, tweets.

1 Introduction

L'édition 2015 du Défi Fouille de Textes est consacrée à l'analyse d'opinion dans un corpus de *tweets*. Cette édition comportait trois tâches dont une découpée en deux sous-tâches : Nous n'avons pas participé à la tâche 3, nous concentrant sur les tâches de classification T1, T2.1 et T2.2 :

T1 Classification des *tweets* selon leur polarité (3 classes) ;

T2 Classification fine des *tweets* ;

- **T2.1** Identification de la classe générique de l'information exprimée dans le *tweet* (4 classes) ;
- **T2.2** Identification de la classe spécifique de l'opinion, sentiment ou émotion (18 classes) ;

T3 Détection de la source, la cible et de l'expression d'opinion.

Les *tweets* ont été annotés manuellement avec les différentes classes auxquels ils se rattachaient. Pour une explication plus précise des modalités d'annotation, nous renvoyons au guide d'annotation mis en ligne par les organisateurs du défi [1]. L'intérêt de ces tâches de classification réside par exemple dans le fait de repérer si un *tweet* a une connotation positive ou négative, s'il est purement informatif ou s'il exprime une opinion...Dans le cas particulier des *tweets*, cela permet d'aller au-delà de la simple description des *tweets* par les *hashtags*. Parmi les débouchés possibles de ce type d'analyse, nous pouvons citer la veille commerciale (popularité d'un produit) et la veille sociétale (viabilité d'un projet politique).

Dans une première approche de la fouille d'opinion, l'indicateur du nombre de *tweets* traitant d'un sujet pourrait suffire à déterminer la popularité de ce sujet. Ceci rappelle un aphorisme célèbre [2] : « Qu'on parle de moi en bien ou en mal, peu importe. L'essentiel, c'est qu'on parle de moi ! ». L'idée est donc que l'on ne s'intéresse qu'au fait que des émotions soient exprimées, quelles qu'elles soient. De telle sorte que l'on accorde une importance centrale à la récurrence d'un thème dans un corpus ou plus généralement dans l'« actualité ». À l'opposé, l'on serait plus indifférent vis-à-vis d'autres thèmes, moins fréquemment rencontrés dans un corpus indépendamment de la polarité de leur traitement. Au contraire, une approche plus « qualitative »s'intéresserait beaucoup plus au contenu réel de ce qui est transmis, à ce qui est dit du sujet. D'un point de vue linguistique, un accent serait donc mis sur le rhème et non plus seulement sur le thème.

Dans la section 2 nous décrivons l'approche que nous avons employée pour cette édition du défi. Dans la section 3 nous présenterons les données mises à disposition pour le défi ainsi que les résultats obtenus par notre approche. Nous proposerons nos conclusions et perspectives de recherche dans la section 4.

2 Description de l'approche

Les travaux en fouille d'opinion sont classiquement répartis entre des approches symboliques (Zhang *et al.*, 2012) et approches dites statistiques (Pak & Paroubek, 2010) avec au centre les approches dites mixtes ou hybrides (Vernier *et al.*, 2009). Notre angle d'attaque pour cette édition du défi est fondé sur la stylométrie, domaine où l'on trouve les trois types d'aproche précités. La stylométrie, parfois nommée *forensic linguistics* dans la littérature, consiste à chercher les indices qui rattachent un énoncé à un style particulier. Ceci peut par exemple permettre de rattacher un texte à un sous-genre, méthode employée dans l'édition 2014 du DEFT pour classer des nouvelles (Lecluze & Lejeune, 2014). Une des applications les plus fréquentes est la tâche d'attribution d'auteur qui consiste à identifier les indices laissés par un auteur dans les textes qu'il a produit de manière à disposer d'un modèle pouvant prédire les auteurs de textes anonymes (Brixtel *et al.*, 2015). (Daelemans, 2013) propose une tripartition des connaissances que l'on peut extraire d'un texte : objectives (informations factuelles), subjectives (opinions exprimées) et meta-connaissances (extraire ce qui n'est pas dans le texte. Dans cette classification, la stylométrie s'attaque donc plus généralement aux méta-connaissances, l'auteur et le genre textuel dans les deux exemples cités.

Notre hypothèse est que la stylométrie peut aussi permettre d'extraire de la connaissance dans le domaine subjectif. Les traits exploités sont des signes de ponctuation et des n-grams de caractères, traits qui sont les plus efficaces selon les travaux récents sur l'attribution d'auteur (Sun *et al.*, 2012).

Pour notre premier *run*, nous avons exploité l'effectif des caractères non-alphanumériques, l'effectif total de lettres et l'effectif total de chiffres dans chaque *tweet*. Pour chacun de ces traits nous avons ajouté en plus de l'effectif, la proportion de cet effectif en fonction de la taille du *tweet*. Pour les *run* 2 et 3 nous avons exploité des n-grams de caractères avec respectivement $n = 1$ et $3 <= n <= 4$. L'objectif du *run* 1 est de mesurer si l'ajout dans les traits de l'effectif de tous les caractères alpha-numériques (et non simplement de la classe lettres et de la classe chiffres) amènerait une perte de résultat. Le *run* 3 correspond à des valeurs de n pour lesquels les modèles en n-grams de caractères sont connus pour être efficaces.

La chaîne de traitement utilisée est illustrée dans la figure 1. Le classifieur utilisé est un SVM (Séparateur à Vaste Marge ou *Support Vector Machine*) à noyau linéaire, classique dans les analyses stylométriques orientées attribution d'auteur. Nous avons utilisé $C = 1$ pour paramètrer la fonction de coût. D'une part, cela permettait de nous placer dans la même configuration que les approches classiques d'attribution d'auteur et donc de faciliter les comparaisons. D'autre part, nous avons observé que faire varier ce paramètre avait une influence marginale sur les résultats obtenus avec les traits utilisés.

1. `https://deft.limsi.fr/2015/guideAnnotation.fr.php?lang=fr` accédé le 28 mai 2015
2. Attribué à Léon Zitrone

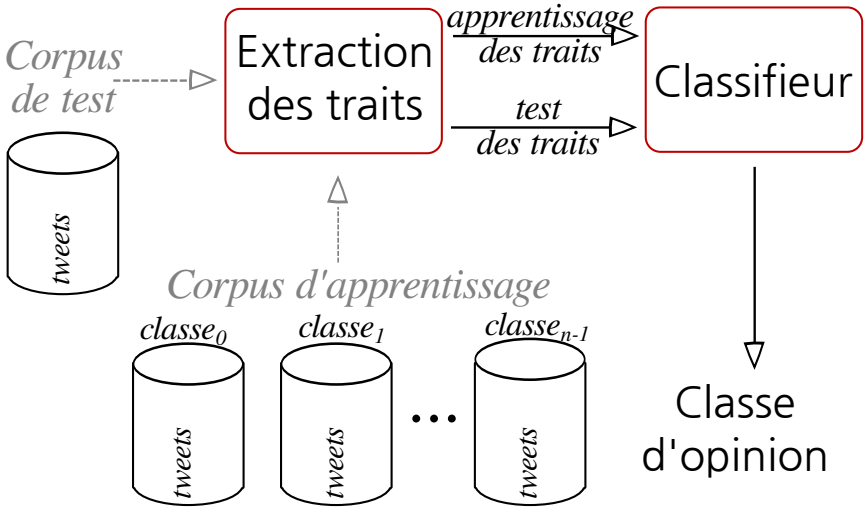

FIGURE 1 – Chaîne de traitement utilisée pour l'apprentissage des classes

3 Résultats

Pour rappel, le *run* 1 correspond à une analyse stylométrique fondée sur l'usage des caractères non alpha-numériques et du nombre de lettres et de chiffres contenus dans chaque *tweet*. La *run* 2 correspond à une classification à partir des 1-grams de caractères tandis que le *run* 3 consiste en une classification à partir de 3-grams et 4-grams de caractères. Nos résultats pendant la phase d'entraînement n'étaient pas très satisfaisants mais nous avons obtenu des résultats bien plus décevants à l'issue de la phase de test (Tableau 1).

	run 1	*run* 2	*run* 3
T1 (3 classes)	0	0,0000986	0,136
T2.1 (4 classes)	0	0	0,097
T2.2 (18 classes)	0	0	0

TABLE 1 – Résultats officiels (macro-précision)

Il s'est avéré que nous avons commis plusieurs erreurs dans la génération des fichers de résultat, aboutissant ainsi dans la moitié des cas à un score nul. Pour cet article, nous avons donc corrigé ces erreurs de manière à présenter des résultats plus fidèles à ce que nous avions constaté lors de la phase d'entraînement. Les résultats recalculés sont présentés dans le tableau 2 et sont, fort heureusement, nettement différents de ceux figurant au classement officiel du défi. Nous souhaitons préciser que les résultats présentés ici sont générés à partir de la même chaîne de traitement que celles que nous avions développé pour le défi. Pour éviter toute confusion, nous avons renommé les runs en « variantes ».

Si nos résultats restent assez décevants, ils amèment tout de même quelques observations intéressantes. La première est que l'approche stylométrique que nous avons utilisé reste inopérante pour la tâche la plus difficile (T2.2). Par contre, elle semble plus prometteuse pour les deux autres tâches bien qu'éloigné des meilleurs résultats des autres équipes participant au défi. Les classes présentes dans ces deux tâches semblent plus compatibles avec l'approche stylométrique que nous

	Résultats globaux du défi			Nos résultats corrigés		
	Moyenne	Médiane	Maximum	Variante 1 (*run 1*)	Variante 2 (*run 2*)	Variante 3 (*run* 3)
T1 (3 classes)	0,581	0,693	0,735	0,369	0,091	0,289
T2.1 (4 classes)	0,408	0,515	0,708	0,318	0,0513	0,251
T2.2 (18 classes)	0,119	0,137	0,231	0,078	0,019	0,059

TABLE 2 – Résultats (macro-précision) après correction des fichiers de sortie

avons adopté. La tâche 2.2 était la plus difficile avec 18 classes différentes mais nous pensions que la méthode stylométrique s'adapterait bien à cette profusion de classes. En effet, il est fréquent en attribution d'auteur de traiter simultanément 50 ou 60 auteurs (et donc autant de classes). Cette hypothès a été contredite par nos résultats. Nous proposons dans la section 4 quelque réfléxions sur ces résultats.

4 Conclusion

Nous avons proposé pour cette édition du Défi Fouille de Textes, une approche stylométrique fondée sur des n-grams de caractères. Notre hypothèse était la suivante : les traits qui permettent d'attribuer la paternité d'un texte devraient être en mesure de déterminer également un style en termes d'opinion ou d'émotion. Indépendamment des problèmes rencontrés dans la phase de test, il apparaît que cette approche était insatisfaisante. Le manque de connaissances proprement linguistiques dans cette méthode a été un facteur trop limitant. Il est également difficile de savoir si ces résultats faibles sont dues à la tâche de fouille d'opinion en elle-même ou bien au genre textuel encore peu formalisé que constituent les *tweets*. Leur faible longueur a pu notamment être un facteur d'échec pour l'approche stylométrique. Par exemple, (Forsyth & Holmes, 1996) considèrent qu'il faut au minimum des messages de 250 mots (1500 caractères) pour que l'analyse stylométrique soit viable. Au contraire, les travaux récents de (Bhargava *et al.*, 2013) et (Almishari *et al.*, 2014) montrent que l'on peut contourner ce problème pour les *tweets* dès lors que l'analyse se porte sur des auteurs très actifs. Malgré toutes ces réserves, il serait intéressant d'explorer de manière plus approfondie ces approches stylométriques pour d'autres tâches de détection d'émotion mais en intégrant, par exemple, des traits plus fins tels que les étiquettes morpho-syntaxiques.

Remerciements

Nous tenons à remercier une fois de plus les organisateurs pour les efforts fournis pour proposer chaque année de nouveaux défis à relever.

Références

ALMISHARI M., KAAFAR D., OGUZ E. & TSUDIK G. (2014). Stylometric linkability of tweets. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, WPES '14, p. 205–208, New York, NY, USA : ACM.

BHARGAVA M., MEHNDIRATTA P. & ASAWA K. (2013). Stylometric analysis for authorship attribution on twitter. In V. BHATNAGAR & S. SRINIVASA, Eds., *Big Data Analytics*, volume 8302 of *Lecture Notes in Computer Science*, p. 37–47. Springer International Publishing.

BRIXTEL R., LECLUZE C. & LEJEUNE G. (2015). Attribution d'Auteur : approche multilingue fondée sur les répétitions maximales. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015)*.

DAELEMANS W. (2013). Explanation in computational stylometry. In A. GELBUKH, Ed., *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, p. 451–462. Springer Berlin Heidelberg.

FORSYTH R. S. & HOLMES D. I. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, **11**(4), 163–174.

LECLUZE C. & LEJEUNE G. (2014). Deft 2014, analyse automatique de textes littéraires et scientifiques en langue française. In *Actes de DEFT 2014 : 10ème DÉfi Fouille de Textes*, p. 11–19, Marseille, France.

PAK A. & PAROUBEK P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

SUN J., YANG Z., LIU S. & WANG P. (2012). Applying stylometric analysis techniques to counter anonymity in cyberspace. *Journal of Networks*, **7**(2).

VERNIER M., MONCEAUX L. & DAILLE B. (2009). DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique. In *Atelier Défi Fouille de Textes (DEFT'09)*, p. 101–112, Paris, France.

ZHANG L., FERRARI S. & ENJALBERT P. (2012). Opinion analysis : The effect of negation on polarity and intensity. In J. JANCSARY, Ed., *Proceedings of KONVENS 2012*, p. 282–290 : ÖGAI. PATHOS 2012 workshop.