CATI : une approche interactive de découverte et de classification de grands corpus de documents

Cédric Boscher, Előd Egyed-Zsigmond, Sylvie Calabretto

Université de Lyon, LIRIS UMR 5205 CNRS
prenom.nom@insa-lyon.fr

Résumé. Dans cet article, nous présentons CATI, une application web interactive d'exploration et de classification de documents. Notre application permet à des utilisateurs non-informaticiens d'explorer et classifier de grandes collections de documents pouvant contenir du texte, des images, et des méta-données telles qu'une date, un auteur, une géolocalisation, etc... CATI fournit un ensemble d'assistants de classification tels qu'un module de détection d'événements, ou encore des méthodes de clustering basées sur des images et du texte. Nous montrons que CATI permet de classifier de grands jeux de données en quelques clics, à l'aide des assistants de classification implémentés et d'assistants permettant à l'utilisateur de sélectionner des attributs méta-données pertinents pour la classification d'un jeu de données.

1 Introduction

Dans un contexte où les réseaux sociaux et la presse en ligne s'imposent comme les principales sources d'informations en ligne permettant d'analyser les tendances de l'opinion publique sur des sujets majeurs d'actualité, les enjeux de méthodes d'extraction d'information appliquées à de grandes collections de documents sont essentiels pour des acteurs de tous types, notamment économiques ou politiques. Diverses approches de fouille de texte combinées à des techniques d'apprentissage automatique permettent de cibler des mots-clés récurrents et des sujets en tendances, lesquels permettent alors de modéliser un ensemble de connaissances et de faciliter la prise de décision d'experts métiers dans un contexte précis. Cependant, ces acteurs n'intègrent pas nécessairement de profils type Data Scientist dans leurs équipes, ou alors ne disposent pas d'employés dédiés à l'annotation massive de données.

Ce papier présente une version améliorée de CATI (Bosetti et al. (2019)), une plate-forme d'assistance à la découverte et à la classification de grandes collections de documents dédiée aux experts métiers et aux utilisateurs non informaticiens. La réalisation d'un tel travail est motivée par l'intérêt potentiel qu'il représente pour différents acteurs économiques ou experts métiers, et répond au besoin d'une solution centrée utilisateur. Elle s'inscrit notamment dans le cadre des projets IDENUM [1] et LIVRONS [2], qui s'intéressent aux représentations visuelles

1. https ://imu.universite-lyon.fr/projet/idenum-identites-numeriques-urbaines/
2. https ://imu.universite-lyon.fr/projet/livrons-livraison-a-velo-representations-sociales-et-donnees-des-reseaux-sociaux-2020s

à grande-échelle de la ville, au moyen d'analyses de publications de photos et de contenus textuels sur les réseaux sociaux.

Dans ce papier, nous présentons :

— Notre proposition d'application interactive de recherche, visualisation et classification de grandes collections de documents, adaptée à des utilisateurs non-informaticiens et conçue pour des documents multi-modaux intégrant du texte, des images et diverses méta-données
— Notre approche de classification assistée, permettant à l'utilisateur de sélectionner ou exclure certains attributs des documents explorés, dépendamment du critère de classification défini par l'utilisateur et du potentiel discriminant des attributs de données analysés
— Une analyse de la qualité de notre approche de classification appliquée à de grandes collections de documents, ainsi qu'une étude d'impact de la définition du critère d'annotation de documents par l'utilisateur et de la prise en compte des méta-données

2 Travaux Connexes

Le marché des solutions d'analyse de documents est dense. Il existe un certain nombre d'outils d'analyse et de classifications de textes, tels que : TXM [3], MonkeyLearn [4], KairnTech [5], Viky.ai [6], GATE [7], IRaMuTeQ [8]. Néanmoins, ces outils ne permettent pas la classification assistée de documents multi-modaux dans un environnement no-code.

La classification de texte est une approche classique de recherche d'informations en lien avec une thématique spécifique au sein d'un corpus de documents issus de réseaux sociaux ; en revanche, le lien avec la thématique n'est pas toujours explicitement déductible de l'information textuelle. Bruijn et al. (2020) s'appuient par exemple sur des méta-données textuelles et temporelles liées à un contexte de publication, afin d'identifier des documents pouvant se référer implicitement à des faits d'inondations. Cette approche s'appuie sur des réseaux de neurones entraînés sur un jeu de données d'informations hydrologiques contextuelles, afin de proposer une classification plus pertinente et basée sur le contexte. Elle montre une amélioration significative de la précision du classifieur entraîné. Les avantages concrets de l'analyse multi-modale pour des documents issus de réseaux sociaux a été mise en exergue par Nikolopoulos et al. (2011), qui montrent l'importance de la sélection d'attributs et de combinaisons optimales pour la classification de données multi-modales, notamment en se basant sur du clustering d'images.

Vijayaraghavan et al. (2017) se concentrent sur l'apprentissage profond multi-tâche et multi-modal pour la classification de profils utilisateurs de réseaux sociaux, en fonction des contenus partagés et des méta-données des contenus. Ils définissent une approche de vectorisation de données multi-modales, ainsi qu'une méthode fondée sur des classifieurs à mécanisme

3. http ://textometrie.ens-lyon.fr
4. https ://monkeylearn.com
5. https ://kairntech.com/fr
6. https ://www.viky.ai
7. https ://gate.ac.uk
8. http ://www.iramuteq.org

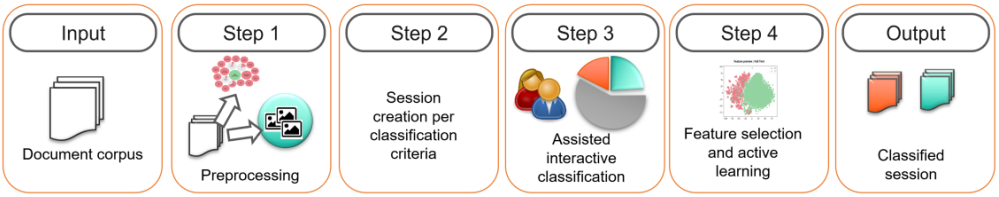

FIG. 1 – *Étapes de classification de corpus dans CATI*

d'attention. Zhou et al. (2020) vont plus loin dans l'analyse multi-modale de documents et présentent SAFE, une méthode d'analyse de similitudes multi-modales apprenant de relations entre textes et images, et quantifie leur similitude d'un point de vue sémantique afin de déterminer si un texte et une image traitent d'un sujet commun, et représentent tous les deux une information pertinente vis à vis d'un sujet particulier.

Considérant l'ensemble des travaux que nous venons de décrire, notre objectif principal dans le cadre de ce travail est de combiner un outil d'annotation de données centré utilisateur non-informaticien et une méthode assistée de classification multi-modale prenant en compte le texte, les méta-données et les images liées à des documents issus de médias en ligne et réseaux sociaux.

3 Description de l'outil

Notre approche consiste à fournir un outil centré utilisateur permettant la classification de grandes collections de documents avec un effort humain réduit, dans un environnement web dit "no-code". Le scénario de classification des corpus est constitué de 4 étapes décrites ci-dessous (voir Fig. 1).

Notre outil, CATI, permet d'importer des corpus de documents multi-modaux comme jeu de données d'entrée. À partir d'un jeu de données, nous créons des sessions virtuelles (Fig. 1, Étape 2) associées à un objectif de classification. Par exemple, nous pouvons définir comme corpus l'ensemble des Tweets contenant le mot clé "Lyon" collectés entre Mars et Septembre 2021, et définir une session dédiée aux tweets liés au football, ainsi qu'une session ayant pour thématique la Fête des Lumières. L'objectif de chaque session est de permettre de classifier les documents du corpus comme positifs ou négatifs, en fonction du critère de classification qui lui est attribué. Si nous reprenons notre premier exemple, les tweets liés au football seront classifiés comme positifs et tous les autres en négatif. D'autres types de corpus peuvent être basés sur des articles de presse, des contenus de blogs, ou tout type de document pouvant contenir des images, des dates, une géo-localisation ou autre méta-données. Pour chaque corpus de documents, CATI applique un ensemble d'étapes de pré-traitement (Fig. 1, Étape 3), décrites dans la section 3.2. Dès la fin du pré-traitement, il est possible de créer autant de sessions que nécessaire. Pour une session donnée, CATI assiste l'utilisateur afin de classifier le plus grand échantillon de documents possibles avec un nombre réduit de clics.

Les assistants de classification implémentés dans CATI sont décrits dans la section 3.3. Ces assistants permettent à la fois d'explorer et d'annoter les documents du corpus.

Après avoir annoté un certain nombre de documents dans une session selon le critère défini , l'utilisateur peut dès lors classifier le reste du corpus avec le module de classification automatique de CATI (Fig. 1, Étape 4). Dans cette étape, nous entraînons un classifieur sur le jeu de données préalablement annoté par l'utilisateur puis nous proposons une classification automatique du reste du corpus, en nous appuyant sur un module de sélection assistée d'attributs multi-modaux à prendre en compte, ainsi qu'un module d'apprentissage actif permettant d'évaluer et améliorer la qualité du modèle de façon assistée.

3.1 Architecture

CATI est une application web basée sur un serveur Python Flask en dorsal, et une interface web basée sur Backbone.JS en frontal. L'application s'interface avec une instance de cluster Elasticsearch (base de données NoSQL orientée documents) utilisée pour l'indexation et la recherche de documents. En parallèle, nous utilisons une application Java de collecte de documents appelant l'API Twitter, permettant d'indexer des Tweets en temps réel et une application C++ pour le clustering d'images quasi-identiques (Gaillard et al. (2018)).

3.2 Méthodes de pré-traitement

L'approche de recherche d'informations dans CATI consiste à agréger des ensembles de documents en nous basant sur des attributs communs tels que des attributs textuels, des attributs visuels lors qu'ils intègrent des images ou encore des attributs temporels, puis à annoter ces documents par lots en quelques clics.

Ces différents attributs sont extraits lors de la phase de pré-traitement :

— **Nettoyage de texte** Nous supprimons les mots vides et procédons à une lemmatisation du texte. Il sera alors possible de nous baser sur la forme radicale des termes que nous voulons rechercher.

— **N-grams.** Pour chaque document, nous extrayons des N-Grams (avec N = 2 ou 3 selon préférence).

— **Clusters d'images** Les documents contenant des images sont regroupés par similarité : images semblables en prenant en compte le bruit, le flou, les niveaux de gris, la rotation ou le recadrage (Gaillard et al. (2018))

— **Détection d'objets dans les images** CATI intègre une approche de reconnaissance d'objets dans les images basée sur des réseaux convolutionnels, YOLO (Redmon et Farhadi (2018)), permettant de reconnaître 80 types d'objets tels que des personnes, des vélos, des bâtiments ou encore des voitures. Il est alors possible de filtrer les documents en fonction des objets détectés (exemple : tous les tweets contenant une image de vélo / cycliste).

— **Détection d'événements.** Nous fournissons une méthode d'agrégation basée sur la détection d'événements, utilisant une version améliorée de MABED (Guille et Favre (2015)). Nous regroupons alors les documents au sein d'événements détectés, basés sur des anomalies de fréquences de termes et sur des regroupements de clusters d'images (Odeh (2018)).

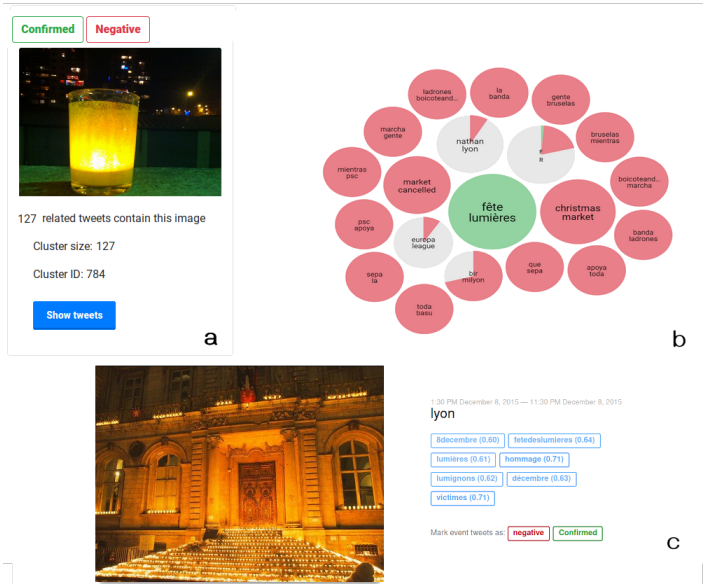

FIG. 2 – *Assistants interactifs d'annotation : Clusters d'images (a), N-Grams (b), Détection d'événements (c)*

3.3 Assistants interactifs d'annotation

CATI intègre un éventail d'assistants d'annotation, permettant l'exploration de documents par lots, basés sur des attributs multi-modaux communs. Il est ensuite possible de classifier ces documents individuellement, ou par lot en un seul clic. L'objectif de cette étape est d'explorer et analyser qualitativement les documents du corpus, mais aussi de constituer un jeu de données d'entraînement pour l'étape de classification, avec un nombre réduit de clics.

La Fig. 2 présente trois assistants se basant sur les attributs extraits lors de la phase de prétraitement. Les clusters d'images (a) contiennent l'ensemble des documents ayant une image similaire. Il est alors possible d'annoter l'ensemble de ces documents en un seul clic, ou d'explorer le lot de documents en détails à l'aide du bouton "Show Tweets". Les N-Grams (b) sont une représentation simplifiée des associations de termes successifs les plus fréquemment détectées dans le corpus. Chaque bulle représente un bi-gram. L'utilisateur peut cliquer dessus pour explorer les documents liés et les annoter (individuellement ou par lot). L'outil de détection d'événements (c) montre les résultats de l'étape de détection d'événements décrite dans la section 3.2. Nous y trouvons notamment les dates d'intervalle de chaque événement, les mots clés les plus significatifs et l'image la plus représentative. Nous affichons également un histogramme de fréquence d'apparition des documents liés à chaque événement, afin d'évaluer leurs cycles de vie respectifs.

L'interface de CATI intègre un moteur de recherche permettant de filtrer les documents. La recherche de documents peut alors s'effectuer sur l'ensemble du corpus, ou alors sur un échantillon de tweets liés à certains mots clés.

FIG. 3 – *Nuage de mot affichant les termes les plus significatifs pour un sous-ensemble de tweets concernant le Football à Lyon*

3.3.1 Termes significatifs et visualisation de tendances

L'interface d'annotation fournit des assistants visuels ayant pour objectif d'aider l'utilisateur à comprendre le champ lexical et les tendances liées à une recherche spécifique dans CATI. Nous fournissons un assistant Nuage de Mots (voir Fig. 3) affichant les termes les plus significativement liés à une requête spécifique (i.e., les termes qui apparaissent, en proportion, plus fréquemment dans le contexte d'une requête que sur l'ensemble du corpus). Chaque terme dispose d'un score évaluant sa significativité dans le contexte de la requête formulée par l'utilisateur. [1] Ce score détermine la taille des mots affichés dans le nuage ; les mots qui apparaissent les plus gros sont les plus significatifs. Chaque mot est cliquable et permet d'afficher et annoter l'ensemble des documents liés.

CATI intègre également un assistant de visualisation chronologique des tendances (voir Fig. 4), qui projette les termes significatifs détectés sur des intervalles de temps réguliers (par exemple, par semaine) et détecte automatiquement les documents considérés les plus représentatifs des tendances sur chaque intervalle. Pour ce faire, nous assignons un score global à chaque document, calculé par la somme pondérée des scores de significativité de chacun de ses termes ; alors, les n documents avec le plus haut score global sont considérés comme les plus représentatifs.

3.4 Classification automatique de corpus

Après avoir annoté un sous-ensemble de documents représentatifs du critère de sélection défini par l'utilisateur, celui-ci peut classifier le reste du corpus via un module de classification automatique assistée.

3.4.1 Sélection d'attributs

Une interface basique de classification permet à l'utilisateur d'entraîner un classifieur sur le sous-ensemble de documents préalablement annotés, et de lancer une classification automatique du reste du corpus en un clic. Dans la configuration standard, le texte des documents est vectorisé (vectoriseur TF-IDF) puis nous entraînons un classifieur SVM linéaire sur les représentations vectorielles des documents annotés, dans le but de classifier automatiquement les documents non annotés.

1. http ://blog.comperiosearch.com/blog/2015/06/10/how-elasticsearch-calculates-significant-terms/

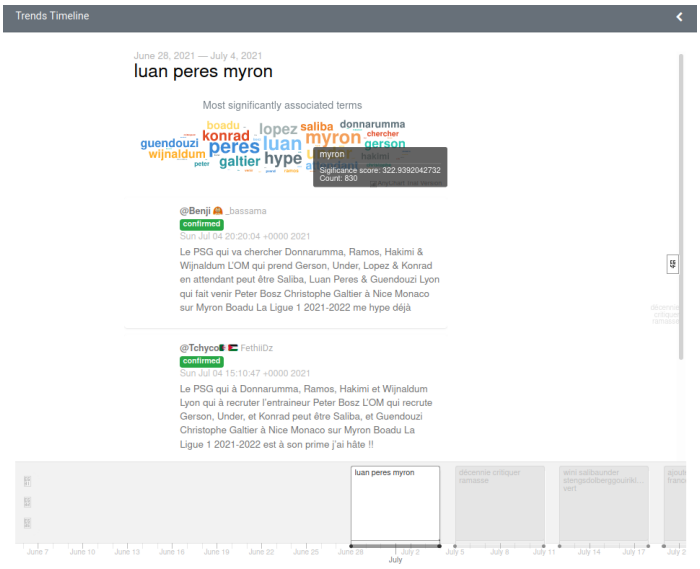

FIG. 4 – *Assistant de visualisation chronologique des tendances, affichant les tendances hebdomadaires pour un sous ensemble de tweets concernant le Football à Lyon*

De façon complémentaire, un mode Classification Experte offre une interface avancée conçue pour configurer une méthode de classification personnalisée ; elle permet de sélectionner les attributs méta-données à prendre en compte pour la classification, le type de vectorisation et le type de classifieur. Un champ de sélection montre les méta-données disponibles pour le corpus sélectionné et l'utilisateur peut sélectionner les plus pertinents, en rapport au critère de classification de la session. Par exemple, le jour de la semaine peut être une information intéressante pour la recherche de documents liés au football, dans la mesure où les matchs de football se déroulement généralement le Mercredi et le Samedi en France. La représentation vectorielle des documents prenant en compte les méta-données, tel que le suggèrent (Vijayaraghavan et al. (2017)), se fait par concaténation de la représentation vectorielle de chaque attribut distinct. Le texte des documents peut être vectorisé à l'aide d'un vectoriseur TF-IDF, d'un vectoriseur Count ou d'un Doc-Embedding fourni par SpaCy Honnibal et al. (2020) . La méthode d'encodage de chaque attribut méta-donnée est ensuite sélectionnée par l'utilisateur. Les attributs catégoriels comme l'auteur, les objets détectés ou le jour de la semaine sont vectorisés à l'aide d'un encodage 1-parmi-n.

Afin d'assister l'utilisateur dans l'analyse du potentiel discriminant de chaque attribut, nous affichons également une projection 2D des données selon chaque attribut, à l'aide d'une réduction de dimension TSNE appliquée à la représentation vectorielle des documents selon l'attribut sélectionné. (voir Fig. 5). Concernant les attributs temporels, nous affichons également un histogramme permettant d'identifier de potentielles dates discriminantes.

La représentation 2-D des données selon chaque attribut affiche les documents positifs en vert et les négatifs en rouge.

FIG. 5 – *Interface de classification experte : Sélection d'attribut et pré-visualisation des représentations vectorielles du texte des documents (Projection 2-D via TSNE)*

3.4.2 Classifieurs

CATI intègre divers classifieurs prenant en entrée des vecteurs avec de grandes dimensions (Machine à vecteurs de support linéaire, Forêt d'arbres de décisions aléatoires, Arbre de Décision, K Plus proches voisins et Régression Logistique). En complément, nous intégrons un classifieur basé sur BERT (Devlin et al. (2019)) et un Perceptron Multi-Couches (MLP à deux couches) (Ostendorff et al. (2019)). Le texte est classifié positivement ou négativement par BERT, puis le Perceptron Multi-couche prend en entrée la concaténation de la sortie de BERT et de la représentation vectorielle des attributs méta-données sélectionnés pour l'entraînement. Nous implémentons également un système de vote entre les 6 classifieurs décrits ; la classe prédite pour chaque document est celle prédite par la majorité des classifieurs.

3.4.3 Apprentissage actif

Afin d'aider l'utilisateur à évaluer et améliorer la précision et le rappel du classifieur entraîné, CATI intègre un module d'apprentissage actif. Après un premier entraînement du classifieur, une sélection de N documents est présentée à l'utilisateur, ainsi que la classe prédite pour chaque document. Nous implémentons différentes stratégies d'apprentissage actif ; les documents peuvent être sélectionnés de façon aléatoire, ou alors en sélectionnant les N documents pour lesquels le score de confiance de prédiction est le plus faible (ou le plus élevé, selon la stratégie adoptée). Ensuite, l'utilisateur peut valider ou corriger les exemples en un clic. Les documents validés sont ajoutés au jeu d'entraînement, le modèle est ré-entraîné et l'interface présente une pré-visualisation des prédictions du classifieur pour les documents non-annotés (sous forme de représentations visuelles permettant de distinguer les positifs et les négatifs). L'utilisateur peut répéter l'étape d'apprentissage actif autant de fois qu'il le juge nécessaire.

TAB. 1 – *Description des jeux de données*

Jeu de données	Nb. Doc.	Nb. Doc. avec images	Nb. positifs	% positifs
FDL 2015	31 006	3 587	1 061	3,42
Lyon Football	1 154 832	53 089	114 778	9,93

4 Évaluation

4.1 Description des jeux de données

L'objectif principal de CATI étant de proposer une manière simple de rechercher et d'extraire des documents pertinents, parmi une grande collection de documents et selon un critère allant du vague au très spécifique (exemples : Football, Fête des Lumières ou activités de cyclologistique), le jeu de données est typiquement censé être séparé en deux classes déséquilibrées : les documents positifs (documents liés au critère défini) et les documents négatifs (contenus divers et variés, non liés au critère défini).

Il existe en réalité certains jeux de données vérité-terrain qui implémentent une telle dichotomie, comme *The Twitter Political Corpus* (Marchetti-Bowick et Chambers (2012)) qui distingue des tweets liés à la politique et des tweets quelconques non liés au sujet (deux classes distinctes), mais ces corpus n'incluent pas d'images, et ne contiennent pas les méta-données des tweets. MVSA (Zhang et al. (2020)) serait un exemple plus représentatif d'un cas d'usage réel, mais il ne contient que du texte et des images.

Pour pallier ces problèmes, nous créons nos propres jeux de données (Table 1) spécifiquement constitués pour notre tache d'évaluation, et basés sur sur des tweets collectés via l'API Twitter.

4.1.1 FDL 2015 :

Nous avons collecté des tweets contenant le mot clé **Lyon** entre **le 7 Décembre 2015 et le 12 Décembre 2015**, puis nous avons classifié les tweets liés à la Fête des Lumière comme positif, le reste comme négatif.

4.1.2 Lyon Football :

Nous avons collecté des tweets contenant le mot clé **Lyon** entre **le 1 Juillet 2021 et le 20 Septembre**, puis nous avons classifié les tweets liés au football comme positifs. Nous nous appuyons sur les mots clés suivants **Foot, Football, PSG, OL, #ArsenalLyon, #OLPSG, Monaco, Atletico**. Le reste du corpus est classifié négativement.

4.2 Classification prenant en compte les méta-données

Afin d'évaluer notre approche de classification automatique de corpus, nous mesurons le F1-Score des modèles en fonction de la taille du jeu d'entraînement (i.e., le nombre de documents à annotés par l'utilisateur avec les assistants d'annotation et notre approche d'apprentissage actif), ainsi que l'impact des attributs méta-données sur la qualité de la classification. Les deux classes étant déséquilibrées, nous mesurons le F1-Score macro moyen.

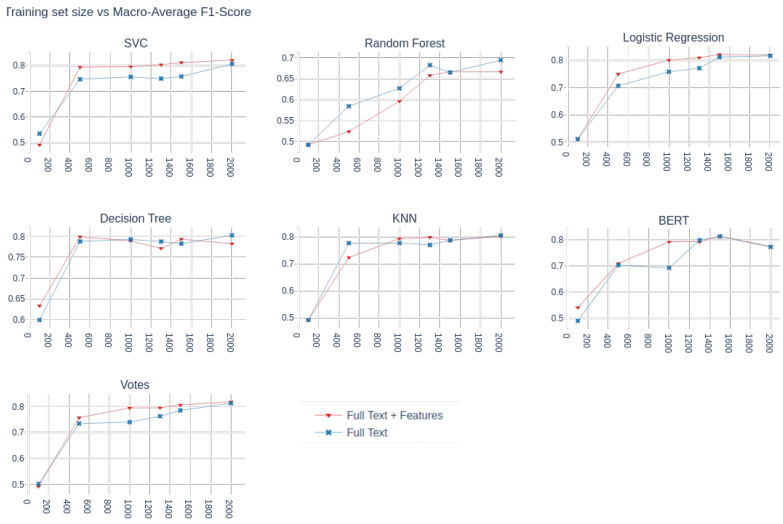

FIG. 6 – *F1-Score des classifieurs pour le jeu de données FDL-2015*

Nous comparons les scores de classification pour nos deux jeux de données selon deux configurations distinctes : texte uniquement, puis texte + deux attributs méta-données (le **Jour de la semaine** et les **Objets détectés dans les images**). Les deux attributs sont encodés en 1-parmi-n, et le texte est vectorisé avec le vectorizer TF-IDF de ScikitLearn. Nous sélectionnons et annotons 50 documents (en préservant la distribution de classes du jeu de données complet), puis nous ajoutons itérativement de nouveaux documents aléatoires au jeu d'entraînement. Pour chaque taille de jeu de données, nous entraînons chaque classifieur et mesurons leur F1-Score respectif.

Les résultats de classification pour **FDL2015** (voir Fig. 6) montre que les courbes avec et sans les attributs méta-données convergent au delà d'un seuil de 2000 documents annotés par l'utilisateur. En revanche, toutes les méthodes de classification implémentées (à l'exception de Random Forest) montrent que le F1-Score est meilleur pour un faible nombre de documents annotés lorsque l'on prend les attributs méta-données en compte. Dans ce cas présent, le système de vote atteint un F1-Score de 81% avec seulement 1000 documents annotés, alors que la même méthode en prenant uniquement le texte en entrée requiert deux fois plus de documents annotés pour atteindre le même score.

Dans le cas de **Lyon Football** (voir Fig. 7), nous observons que le classifieur Arbre de Décision est la meilleure approche, avec un pic de F1-Score de 95,11% pour un jeu d'entraînement de 1000 documents. Pour les autres classifieurs (à l'exception de BERT et KNN), le F1-Score augmente très rapidement au delà de 90% avec un jeu d'entraînement relativement réduit (1000 documents), mais contrairement au cas du jeu de données FDL2015, l'inclusion des méta-données n'apporte pas d'amélioration significative du modèle. Le texte constitue déjà, en soi, une information fortement discriminante.

FIG. 7 – *F1-Score des classifieurs pour le jeu de données Lyon Football*

5 Conclusion et Travaux Futurs

Dans ce papier, nous avons présenté un outil de classification interactif pour de grands corpus de documents, prenant en compte des données multi-modales : texte brut, images et méta-données. Nous avons élaboré une interface centrée-utilisateur permettant de visualiser des tendances et des documents représentatifs selon un certain critère parmi un grand corpus, et permettant d'annoter de grandes quantités de documents agrégés selon diverses similarités, en seulement quelques clics. Nous avons montré, dans notre cas, qu'il est possible de classifier un grand corpus (plus d'1M de documents) avec des classes très déséquilibrées, en fournissant un jeu d'entraînement de l'ordre du millier de documents annotés. Nous avons montré que lorsque le critère de classification est complexe, l'inclusion des méta-données dans le processus de classification peut améliorer la qualité des modèles et réduit considérablement la taille du jeu de données d'entraînement à annoter.

Dans nos travaux futurs, nous évaluerons plus qualitativement la valeur ajoutée de l'outil en terme de prise en main par des utilisateurs non-informaticiens ; celui-ci est activement pris en main par des partenaires académiques (Économie des transports) dans le cadre du projet LIVRONS [9] au moment où nous écrivons ce papier, ce qui constitue pour nous un scénario type d'usage de CATI par un public non-informaticien.

Références

Bosetti, G., E. Egyed-Zsigmond, et L. Ono (2019). CATI : An Active Learning System for Event Detection on Mibroblogs' Large Datasets :. In *Proceedings of the 15th International Conference on Web Information Systems and Technologies*, Vienna, Austria, pp. 151–160. SCITEPRESS - Science and Technology Publications.

9. https ://imu.universite-lyon.fr/projet/livrons-livraison-a-velo-representations-sociales-et-donnees-des-reseaux-sociaux-2020/

Bruijn, J. A. d., H. d. Moel, A. H. Weerts, M. C. d. Ruiter, E. Basar, D. Eilander, et J. C. J. H. Aerts (2020). Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network. *Computers & Geosciences 140*, 104485.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics.

Gaillard, M., E. Egyed-Zsigmond, et M. Granitzer (2018). CNN features for Reverse Image Search. *Document Numérique 21*(1-2), 63–90.

Guille, A. et C. Favre (2015). Event detection, tracking, and visualization in Twitter : a mention-anomaly-based approach. *Social Network Analysis and Mining 5*(1), 18. arXiv : 1505.05657.

Honnibal, M., I. Montani, S. Van Landeghem, et A. Boyd (2020). spaCy : Industrial-strength Natural Language Processing in Python.

Marchetti-Bowick, M. et N. Chambers (2012). Learning for microblogs with distant supervision : Political forecasting with Twitter. pp. 603–612.

Nikolopoulos, S., E. Giannakidou, I. Kompatsiaris, I. Patras, et A. Vakali (2011). *Combining Multi-modal Features for Social Media Analysis*, pp. 71–96.

Odeh, F. (2018). Event detection in heterogeneous data streams. Technical report, Lyon. pp. 28.

Ostendorff, M., P. Bourgonje, M. Berger, J. M. Schneider, G. Rehm, et B. Gipp (2019). Enriching bert with knowledge graph embeddings for document classification.

Redmon, J. et A. Farhadi (2018). Yolov3 : An incremental improvement.

Vijayaraghavan, P., S. Vosoughi, et D. Roy (2017). Twitter Demographic Classification Using Deep Multi-modal Multi-task Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Vancouver, Canada, pp. 478–483. Association for Computational Linguistics.

Zhang, K., Y. Geng, J. Zhao, J. Liu, et W. Li (2020). Sentiment analysis of social media via multimodal feature fusion. *Symmetry 12*(12).

Zhou, X., J. Wu, et R. Zafarani (2020). SAFE : similarity-aware multi-modal fake news detection. *CoRR abs/2003.04981*.

Summary

In this paper we present CATI, an approach for multi modal document classification implemented on an assisted, interactive document collection manipulation web application. The application helps non computer scientist users to discover, browse and classify large document collections, where documents contain text and can come with images and metadata such as timestamp, author, geolocation, etc. CATI provides classification assistants such as event detection, text and image based document clustering. It comes with an interface that helps users select among several text and other information based features to classify the documents. Our study shows that using the classification assistants and helping users choose the right features gives good classification results for large document collection within a few clicks.