Classification automatique d'articles encyclopédiques

Ludovic Moncla*, Khaled Chabane*, Alice Brenon*,**

*INSA Lyon, LIRIS UMR CNRS 5205, Lyon, France
prenom.nom@insa-lyon.fr,
**ICAR UMR CNRS 5191, Lyon, France

Résumé. Cet article propose une étude comparative de différentes approches de classification supervisée appliquées à la classification automatique d'articles encyclopédiques. Notre corpus d'apprentissage est constitué des 17 volumes de texte de l'Encyclopédie de Diderot et d'Alembert (1751-1772) représentant un total d'environ 70 000 articles. Nous avons expérimenté différentes approches de vectorisation de textes (sac de mots et plongement de mots) combinées à des méthodes d'apprentissage automatique classiques, d'apprentissage profond et des architectures BERT. En plus de la comparaison de ces différentes approches, notre objectif est d'identifier de manière automatique les domaines des articles non classés de l'Encyclopédie (environ 2 400 articles). Le meilleur modèle permet d'obtenir 89% de f-mesure moyenne pour l'ensemble des 38 classes. Par ailleurs, notre étude met en avant la difficulté à distinguer certaines classes proches sémantiquement. L'ensemble du code développé ainsi que les résultats obtenus dans le cadre de ce projet sont disponibles en open-source [1].

1 Introduction

Dans cet article, nous présentons une étude comparative de méthodes d'apprentissage supervisé pour la classification d'articles encyclopédiques. Nous nous sommes en particulier intéressés à l'*Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres (1751-1772)* de Diderot et d'Alembert. Ce jeu de données (corpus OCRisé) est proposé par deux organismes différents : l'ENCCRE [2] (Édition Numérique Collaborative et CRitique de l'Encyclopédie) et l'ARTFL [3] (*American and French Research on the Treasury of the French Language*). L'édition numérique de l'Encyclopédie proposée par l'ENCCRE (Guilbaud, 2017) est conçue pour être enrichie de façon collaborative autour d'un ouvrage original conservé à la Bibliothèque Mazarine. La plateforme en ligne permet de consulter à la fois la version originale et la transcription des articles et donne également accès aux notes, commentaires et notices rédigées par les éditeurs-annotateurs de l'ENCCRE. De la même manière, l'ARTFL propose un accès aux différents volumes OCRisés de l'Encyclopédie grâce à sa plateforme Philologic [4]. Du point de vue de la structure, les articles de l'Encyclo-

1. https://gitlab.liris.cnrs.fr/geode/EDdA-Classification
2. http://enccre.academie-sciences.fr/encyclopedie/
3. https://artfl-project.uchicago.edu/
4. https://artfl-project.uchicago.edu/philologic4

pédie se composent de différents éléments tels que la vedette, le désignant, des indications grammaticales, une signature, etc. De manière systématique, la notion d'article renvoie à la notion de vedette, il s'agit d'un mot ou groupe de mots (en capitale) qui se situe en tête d'article et qui en définit le titre et le sujet. Par exemple, l'article présenté en figure 1 a pour vedette le mot "EVIAN". Le désignant fait référence au mot (ou groupe de mots) figurant après la vedette, souvent entre parenthèses, et permet d'indiquer le domaine (ou champ de connaissance) auquel appartient l'article. Il existe une multitude de formulations très variées et non normalisées de ces désignants, plus de 7 000 formes uniques. Afin de réduire ce nombre l'ENCCRE a procédé à plusieurs opérations pour produire des "désignants explicités" telles que la suppression des variations typographiques, la suppression des formules d'introduction, la suppression des abbréviations, la modernisation de l'orthographe et l'uniformisation des formules proches. À la suite de ces opérations, il résulte 2 160 désignants explicités. Ce nombre restant trop élevé pour constituer une typologie de domaines utilisable dans une interface de recherche, une autre étape a alors consisté à construire une liste de domaines. L'ENCCRE propose ainsi 327 domaines qu'ils ont par la suite regroupés en 44 ensembles de domaines. Pour l'article *EVIAN*[5] (voir figure 1) le désignant *(Géog. mod.)* a été explicité "Géographie ancienne" et rattaché au domaine "Géographie". La difficulté au-delà de normaliser et réduire le nombre de classes est que tous les articles ne contiennent pas de désignants. Ainsi, 12 635 articles ne contiennent pas de désignants et ce nombre est réduit à 2 392 après une étape de correction manuelle qui a consisté à assigner des désignants implicites récupérables par rapport au contexte et au contenu des articles.

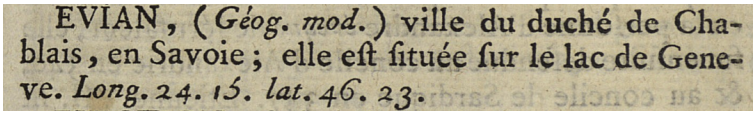
EVIAN, (*Géog. mod.*) ville du duché de Chablais, en Savoie; elle eſt ſituée ſur le lac de Geneve. *Long.* 24. 15. *lat.* 46. 23.

FIG. 1: Article "Evian", vol. VI, n° 417.

L'ARTFL de son côté ne propose pas de notes critiques ou de commentaires des articles ni de regroupement ou classification des désignants en domaines ou ensemble de domaines. Le traitement réalisé s'est concentré sur la normalisation des désignants en *classes* sans (ou avec très peu) d'intervention humaine. Ainsi pour l'article EVIAN (figure 1), l'ARTFL l'associe simplement à la classe normalisée *Géographie moderne*[6]. Sur les 77 085 articles disponibles dans la version proposée par l'ARTFL, 55 248 sont associés à une classe et on retrouve 1 654 classes (normalisées). Il y a ainsi 21 837 articles non classés. Certains travaux menés par les équipes de l'ARTFL se sont intéressés à la classification automatique de ces articles non-classés. Horton et al. (2009) ont expérimenté la méthode de classification naïve bayésienne pour deux tâches, d'une part associer une classe aux articles non classés et d'autre part appliquer le modèle sur les articles déjà classés pour comparer les résultats mais aussi pour déterminer quels mots ont le plus d'importance pour la prise de décision. Dans ce travail, les auteurs ne proposent pas d'évaluation chiffrée des performances du modèle entraîné mais se sont intéressés à une évaluation ciblée sur un petit échantillon d'articles. Plus récemment, Roe et al. (2016) ont proposé l'utilisation de la méthode LDA (*Latent Dirichlet Allocation*) pour

5. http://enccre.academie-sciences.fr/encyclopedie/article/v6-188-0/
6. https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/6/421/

l'analyse du discours encyclopédique du XVIIIème siècle. L'objectif est d'aller au-delà de la classification faite par les auteurs et éditeurs de l'Encyclopédie et d'observer les recoupements thématiques (ou discursifs) entre les classes.

Cet article est structuré de la façon suivante, la section 2 décrit la méthodologie proposée pour l'étude des approches de classification automatique appliquées à la classification d'articles encyclopédiques. La section 3 présente les différentes expérimentations ainsi que les résultats obtenus. Enfin, la section 4 conclut cet article et présente les limites et perspectives de notre étude.

2 Méthodologie

2.1 Préparation du corpus

Dans ce travail, nous exploitons le corpus OCRisé de l'Encyclopédie fourni par l'ARTFL et nous nous concentrons uniquement sur les 17 volumes de texte (77 085 articles) fournis au format XML-TEI. Chaque fichier correspond à un article et comprend en plus du corps de l'article un certain nombre de métadonnées telles que la vedette, le ou les auteurs, le volume, le numéro, le désignant (lorsqu'il est exprimé) et la classe normalisée (lorsque le désignant existe).

La première étape du travail consiste à extraire les données des fichiers XML-TEI et les prétraiter afin d'être exploitables par les algorithmes de classification supervisée. Les algorithmes vont utiliser deux entrées, le texte des articles et leurs classes correspondantes. Comme nous l'avons décrit en introduction, le nombre très important de désignants (plus de 7 000 occurrences) et de classes normalisées (1 654) répertoriés par l'ARTFL représente un frein pour l'entraînement efficace de modèles de classification supervisée. Notre étude du corpus a montré que le nombre important de classes entraîne un fort déséquilibre entre les classes avec quelques classes très peuplées et un nombre très important de classes peu peuplées. On re-

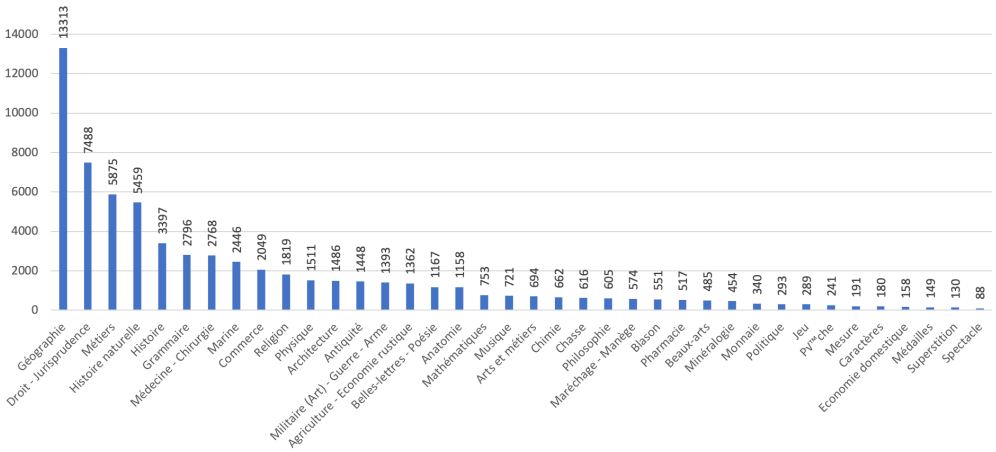

FIG. 2: Nombre d'articles par classe (ensembles de domaines).

trouve par exemple, 1 154 classes avec moins de 5 articles, seulement 122 classes avec 50 articles ou plus et 8 classes avec plus de 1 500 articles. Pour cette raison, nous avons choisi d'utiliser comme classes les ensembles de domaines définis par l'ENCCRE. Nous avons par ailleurs réduit le nombre de 44 ensembles de domaines à 38 en regroupant ceux qui concernent les métiers (tels que *Métiers de l'alimentation*, *Métiers du bois*, *Métiers du cuir et des peaux*, *Métiers du métal, du minéral et dérivés*, etc.). Cependant malgré un nombre de classes réduit, la répartition des instances dans les classes n'est pas équilibrée (figure 2). La classe "Géographie" par exemple comprend 13 313 instances alors que la deuxième classe "Droit-Jurisprudence" en comprend seulement 7 488, et certaines classes comprennent un nombre d'instances inférieur à 1 000 comme "Mathématiques", "Musique" et "Arts et métiers". Pour réduire le biais pouvant être introduit par cet important déséquilibre entre les classes, nous avons rééquilibré en procédant à un échantillonnage des classes sur-représentées. En fixant un seuil maximum du nombre d'articles par classe, nous limitons l'écart entre les classes les moins peuplées et les plus peuplées. Ce seuil a été fixé de manière expérimentale et est décrit dans la partie expérimentation (section 3).

Par ailleurs, comme nous l'avons également indiqué en introduction, un certain nombre d'articles n'ont pas de classe (ou d'ensemble de domaines) associée, ce qui nous a conduit à découper le jeu de données en deux, d'une part un jeu labellisé pour l'entraînement et l'évaluation et d'autre part un jeu non labellisé. Une fois le jeu d'entraînement et de validation constitué (à partir du jeu labellisé) où chaque article est associé à une (ou plusieurs) classe et après suppression des valeurs nulles ou aberrantes, nous avons pu procéder au pré-traitement du texte. Cette étape de pré-traitement se compose de plusieurs sous-tâches telles que la suppression d'articles trop courts (seuil fixé à 15 mots), la suppression des désignants, la tokenisation, la lemmatisation et la suppression des mots vides ou trop fréquents.

2.2 Vectorisation des textes

Notre objectif dans ce travail est de mener une étude comparative des différentes approches de classification. Cela inclut l'étape de vectorisation (ou extraction des caractéristiques) ainsi que sa combinaison avec différents algorithmes de classification. Il s'agit d'une étape classique en traitement automatique des langues qui permet de transformer les textes en vecteurs de valeurs numériques. Dans cette étude, nous nous sommes en particulier intéressés à l'approche de représentation en sacs de mots, ainsi qu'aux plongements de mots statiques et dynamiques.

La première approche consiste à considérer les documents comme des sacs de mots (Salton et McGill, 1986). L'ensemble des termes (tokens ou lemmes) constitue le vocabulaire et représente les dimensions des vecteurs associés à chaque document. Pour chaque dimension, les valeurs peuvent être égales ou supérieures à 0 selon la fréquence d'apparition du terme dans le document. Cette approche produit une représentation vectorielle de très grande dimension et très creuse car chaque article ne contient qu'une faible quantité des termes présents dans l'ensemble du vocabulaire. La représentation en sacs de mots a connu de nombreuses évolutions comme par exemple l'approche TF-IDF (*Term Frequency–Inverse Document Frequency* qui introduit une pondération des termes en fonction de leur fréquence d'apparition dans l'ensemble du corpus mais également au sein d'un même document. De la même manière que pour le sac de mots, cette méthode produit une représentation vectorielle de très grande taille et très creuse.

Les méthodes de plongement de mots ou plongement lexical (*embeddings*) capturent le contexte des mots au sein de l'espace vectoriel. Ces représentations sont pré-entrainées sur la base de ressources textuelles très volumineuses par apprentissage auto-supervisé grâce à un réseau de neurones. Il existe deux architectures permettant d'entraîner un plongement de mots : CBOW et skip-gram. La première a pour tâche de prédire un mot en fonction de son contexte, inversement la méthode skip-gram vise à prédire les mots du contexte. Contrairement aux méthodes "sac de mots", les plongements de mots produisent des vecteurs denses et de faible dimension. Cependant, les méthodes telles que Word2Vec (Mikolov et al., 2013) sont peu adaptées pour la représentation de phrases ou de textes longs (plus de 10 à 15 mots). Pour cette raison, lors de nos expérimentations nous avons utilisé la méthode Doc2Vec (Le et Mikolov, 2014) basée sur Word2Vec mais adaptée pour la vectorisation de documents.

Les plongements de mots tels que Word2Vec produisent une représentation vectorielle unique, combinant les différents contextes d'un mot au sein d'un même vecteur. D'autres méthodes plus récentes, telles que BERT, proposent des plongements de mots dynamiques où chaque mot possède une représentation en fonction de son contexte d'apparition dans la phrase. BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2018) utilise des réseaux de neurones de type *transformer* (Vaswani et al., 2017) et la notion de masque pour prédire des mots ou la phrase suivante. En plus de BERT, nous avons également utilisé CamemBERT (Martin et al., 2019) qui est une adaptation pré-entraînée exclusivement sur des documents en langue française.

2.3 Classification supervisée

L'étape de classification supervisée prend en entrée la représentation vectorielle des articles du jeu d'entraînement ainsi que la classe associée à chaque article.

Dans cette étude, nous avons testé différents algorithmes de classification : des méthodes d'apprentissage automatique classiques et des méthodes d'apprentissage profond utilisant différentes architectures de réseaux de neurones. Concernant les approches classiques, nous avons testé la méthode naïve bayésienne (*Naive Bayes*), la régression logistique (*Logistic regression*), les forêts aléatoires (*Random forest)* et la machine à vecteurs de support (SVM, *Support Vector Machine*). La méthode naïve bayésienne n'est utilisable qu'avec les représentations vectorielles de type sacs de mots (sac de mots et TF-IDF) car les vecteurs produits par ces méthodes ne contiennent pas de valeurs négatives contrairement aux plongement de mots. Nous avons également expérimenté une descente de gradient stochastique (SGD, *Stochastic Gradient Descent)* comme méthode d'optimisation combinée à une machine à vecteurs de support. Le SGD a la particularité de mieux s'adapter aux données de grandes dimensions et peu denses comme c'est le cas pour les représentations vectorielles en sacs de mots.

En plus des méthodes d'apprentissage classiques, nous avons également expérimenté des méthodes d'apprentissage profond telles que les CNN (*Convolutional Neural Network*) et LSTM (*Long Short Term Memory*). Les réseaux convolutionnels se composent de deux opérations : convolution et *max-pooling*. La convolution a pour rôle d'extraire les propriétés des données et le *max-pooling* sert à compresser les résultats de la convolution. Les LSTM sont basés sur des modèles neuronaux récurrents (RNN) et sont particulièrement adaptés pour le traitement de séquences de données. Ils utilisent des blocs de mémoire intermédiaire entre chaque couche pour capturer des dépendances à plus longue portée que les modèles neuronaux classiques. Lors de l'entraînement de modèles de classification par apprentissage profond,

deux approches sont possibles pour l'étape de vectorisation. La première consiste à entraîner un plongement de mots dont les poids de départ sont fixés de manière aléatoires alors que la deuxième consiste à utiliser un plongement de mots déjà pré-entrainé. Dans les deux cas, il s'agit d'ajouter une couche *embedding* au réseau de neurones. Pour nos expérimentations nous avons utiliser la librairie Keras pour les CNN et LSTM et le modèle FastText (Bojanowski et al., 2017) pour la couche embedding pré-entrainée. Enfin, en complément des plongements de mots statiques tels que Doc2Vec ou FastText, les modèles de langue dynamiques tels que BERT peuvent être ré-entraînés sur une tâche ou un corpus spécifique, on parle alors de *fine-tuning*. Cette architecture de type *transformer* obtient les meilleurs résultats de l'état de l'art sur de nombreuses tâches en traitement automatique des langues telles que la classification de document. Une limite de BERT est la longueur maximale de la séquence de tokens que cette architecture peut prendre en entrée (512).

3 Expérimentations

Nos expérimentations concernent l'étude de différentes approches de classification comprenant deux étapes principales : la vectorisation et la classification supervisée. Nous avons testé et comparé les différentes combinaisons suivantes :

1. vectorisation en sac de mots et apprentissage automatique classique (Naive Bayes, Logistic regression, Random Forest, SVM et SGD) ;
2. vectorisation en plongement de mots statiques (Doc2Vec) et apprentissage automatique classique (Logistic regression, Random Forest, SVM et SGD) ;
3. vectorisation en plongement de mots statiques (FastText) et apprentissage profond (CNN et LSTM) ;
4. approche *end-to-end* utilisant un modèle de langue pré-entraîné (BERT, CamemBERT) et une technique de *fine-tuning* pour adapter le modèle sur notre tâche de classification.

Pour les algorithmes de classification classiques, nous utilisons la librairie Scikit-learn [7] et la méthode *GridSearchCV()* pour régler les hyper-paramètres. Pour les réseaux neuronaux nous avons fixé une taille de lot de 256, une taille de vecteur de 300 (pour la couche embeddings) et 20 époques (*epochs*). Nous utilisons une fonction d'activation de type Relu en entrée et *softmax* en sortie, un taux d'abandon de 20% (*dropout*), un optimiseur de calcul du gradient de type Adam et une fonction d'erreur de type entropie croisée. Pour les modèles de type BERT, nous avons fixé une taille de lot de 8 (pour des problèmes de mémoire) et 4 époques.

3.1 Description des jeux de données

Pour l'entraînement et l'évaluation de nos modèles, nous utilisons trois jeux de données (entraînement, validation et test). Ces trois jeux (Table 1) sont constitués des articles catégorisés de l'Encyclopédie issus de deux sources : l'ARTFL et l'ENCCRE. Les jeux d'entraînement et de validation se composent des articles classés par les deux sources alors que le jeu de test se compose uniquement des articles classés par l'ENCCRE et non classés par l'ARTFL (*unclassified*). Les articles non classés, par aucune des deux sources, ne sont ni utilisés pour

7. https://scikit-learn.org/stable/

l'entraînement ni pour l'évaluation mais sont mis de côté et seront classés par la suite grâce au meilleur modèle généré.

Jeux de données	# articles
corpus complet	74 190
corpus nettoyé	54 734
entraînement (complet)	30 650
validation (complet)	10 947
entraînement (seuil max 500)	12 186
validation (seuil max 500)	7 369
entraînement (seuil max 1 500)	21 129
validation (seuil max 1 500)	10 079
test	13 137

TAB. 1: Nombre d'articles pour les différents jeux de données.

A partir des articles fournis au format XML-TEI par l'ARTFL, nous avons procédé à une première étape permettant d'y ajouter la classification fournie par l'ENCCRE (ensembles de domaines). Dans cette phase du projet, nous ne prenons en compte qu'un seul ensemble de domaines par article. En effet, certains articles (3 654, soit 5%) sont associés à plusieurs ensembles de domaines, pour ceux-là nous ne conservons que le domaine principal. Après une étape de nettoyage, comprenant la suppression des articles non classés et des valeurs nulles ou aberrantes (11 619 articles, soit 17%) mais également des articles ayant moins de 15 mots (7 837, soit 12%), nous obtenons 54 734 articles classés pour 38 ensembles de domaines. Nous avons pris 20% des données pour constituer le jeu de test, le reste étant utilisé pour l'entraînement et la validation. Les classes étant très déséquilibrées (fig. 2), nous avons souhaité évaluer l'impact d'un échantillonnage des classes les plus peuplées afin de réduire leur écart avec les classes les moins peuplées. Nous avons comparé les résultats pour trois valeurs de seuil maximum fixées expérimentalement (nous avons choisi les seuils de 500, de 1 500 et sans limite d'articles maximum par classe). Nous obtenons ainsi 19 555 articles pour le seuil de 500, 31 208 articles avec le seuil de 1 500 et 41 597 sans limite. Enfin, nous avons procédé au découpage de ces trois ensembles d'articles en deux sous-jeux de données (entraînement et validation) selon une répartition 70/30 (Table 1) et où chaque classe est représentée de manière équivalente entre les deux jeux.

3.2 Résultats

Pour l'évaluation de la classification, nous utilisons les mesures classiques : précision, rappel et f-mesure, et afin d'obtenir un indicateur unique pour l'ensemble des classes, nous utilisons également une moyenne des scores obtenus (*weighted f1-score*) pour les 38 ensembles de domaines que l'on cherche à catégoriser. Le tableau 2 présente les f-mesures moyennes obtenues pour les différentes combinaisons testées (vectorisation + classification). On observe que les résultats sont quasi identiques pour les jeux de validation et de test. La méthode *Random Forest* obtient les moins bons résultats quelle que soit la méthode de vectorisation ou l'échantillonnage (entre 28% et 50%). La méthode *Naive Bayes* obtient des résultats entre 37% et 74% avec un impact très significatif de l'échantillonnage pour la vectorisation TF-IDF. Les méthodes *Logistic regression*, *SGD* et *SVM* obtiennent des résultats très proches et

Classifieur	Vectorisation	Validation			Test		
		(1)	(2)	(3)	(1)	(2)	(3)
Naive Bayes	Bag of Words	0.71	0.68	0.61	0.72	0.68	0.61
	TF-IDF	0.74	0.59	0.37	0.74	0.59	0.37
Logistic Regression	Bag of Words	0.85	0.85	0.86	0.85	0.85	0.86
	TF-IDF	0.88	0.88	0.89	0.88	0.88	0.88
	Doc2Vec	0.38	0.38	0.44	0.39	0.39	0.44
Random Forest	Bag of Words	0.50	0.49	0.16	0.50	0.49	0.17
	TF-IDF	0.48	0.48	0.15	0.48	0.48	0.16
	Doc2Vec	0.28	0.29	0.37	0.28	0.29	0.37
SGD	Bag of Words	0.85	0.86	0.86	0.85	0.86	0.86
	TF-IDF	0.88	0.88	0.88	0.88	0.88	0.88
	Doc2Vec	0.42	0.41	0.43	0.43	0.42	0.44
SVM	Bag of Words	0.86	0.86	0.88	0.85	0.85	0.88
	TF-IDF	0.87	0.87	0.87	0.86	0.86	0.87
	Doc2Vec	0.31	0.31	0.42	0.32	0.32	0.43
CNN	FastText	0.04	0.05	0.09	0.04	0.05	0.09
LSTM		0.10	0.10	0.12	0.10	0.10	0.12
BERT Multilingual (*fine-tuning*)	-	0.84	0.88	0.89	0.84	0.88	0.89
CamemBERT (*fine-tuning*)	-	0.82	0.86	0.88	0.82	0.86	0.88

TAB. 2: F-mesures moyennes des différents modèles pour les jeux de validation et de test avec un échantillonnage max de 500 (1), 1 500 (2) et sans échantillonnage (3).

les meilleurs sont ceux associés à une vectorisation TF-IDF. La représentation en plongement de mots Doc2Vec produit des résultats décevants et bien en-dessous des représentations en sac de mots. On note que les scores augmentent en fonction de la taille de l'échantillonnage, cela permet de faire l'hypothèse que cette approche nécessite un plus grand jeu de données pour être performante. Les approches d'apprentissage profond utilisant des réseaux neuronaux (CNN et LSTM) obtiennent des scores très faibles (entre 4% et 12%). Par rapport aux résultats obtenus dans la littérature, ceci peut s'expliquer par la faible quantité de données disponible. L'échantillonnage semble avoir un impact mais les résultats sont trop faibles pour en tirer des conclusions. Comme on le voit dans le tableau 3 les seules classes qui obtiennent un score supérieur à 10% sont les deux classes les plus peuplées. Le ré-entraînement des modèles de langue BERT Multilingual et CamemBERT sur notre tâche de classification obtient des résultats presque identiques aux méthodes de classification classiques et permet d'atteindre au mieux 89% de f-mesure. Les résultats entre les deux modèles (BERT Multilingual et CamemBERT) sont très proches en terme de moyenne globale mais diffèrent selon les classes. Par exemple, la classe *Mesure* obtient 74% avec BERT contre 0% avec CamemBERT, la classe *Minéralogie* 65% contre 0%, la classe *Arts et métiers* 51% contre 23%, la classe *Economie domestique* 58% contre 76% et la classe *Superstition* 73% contre 69%. Les plus grands écarts étant pour des classes avec un faible support, l'impact sur la moyenne globale est faible et est compensé par de plus faibles écarts sur les classes les plus peuplées. Afin de permettre une analyse comparative plus fine de nos expérimentations, l'ensemble des résultats est accessible en ligne [8]. Des expérimentations menées sur la taille du texte donné en entrée de l'apprentissage montrent que les résultats sont similaires selon qu'on considère les textes complets des articles ou seulement leur premier paragraphe. Cela signifie que le lexique utilisé dans les pre-

8. https://gitlab.liris.cnrs.fr/geode/EDdA-Classification

mières phrases d'un article contient toute l'information dont le modèle a besoin pour identifier le domaine auquel cet article appartient. De manière générale, on observe qu'augmenter la taille de l'échantillon (nombre d'articles maximum par classe) permet d'améliorer les résultats (à l'exception de la méthode *Random Forest* et de la combinaison *Naive Bayes* + TF-IDF), cependant cela accroît l'écart entre les classes les moins peuplées et les plus peuplées. Cela permet donc de gagner quelques points sur la moyenne globale mais en fait perdre pour les classes sous-représentées.

Ensemble de domaines	#	(1)	(2)	(3)	Ensemble de domaines	#	(1)	(2)	(3)
Géographie	2 870	0.98	0.22	0.99	Arts et métiers	132	0.45	0.00	0.51
Droit - Jurisprudence	1 452	0.92	0.39	0.94	Blason	126	0.93	0.00	0.93
Métiers	1 220	0.87	0.07	0.89	Chasse	124	0.92	0.01	0.92
Histoire naturelle	1 130	0.92	0.06	0.95	Maréchage [...]	118	0.90	0.00	0.88
Histoire	726	0.76	0.08	0.80	Chimie	115	0.75	0.02	0.72
Grammaire	575	0.77	0.08	0.81	Philosophie	115	0.75	0.01	0.69
Médecine [...]	535	0.87	0.07	0.87	Beaux-arts	103	0.86	0.00	0.84
Marine	454	0.93	0.03	0.94	Monnaie	74	0.81	0.00	0.79
Commerce	437	0.85	0.04	0.85	Pharmacie	75	0.65	0.00	0.58
Religion	389	0.89	0.02	0.90	Jeu	67	0.85	0.00	0.87
Architecture	326	0.88	0.01	0.88	Pêche	48	0.93	0.00	0.90
Antiquité	321	0.80	0.01	0.82	Mesure	43	0.65	0.00	0.74
Physique	309	0.85	0.04	0.86	Economie domestique	31	0.75	0.00	0.58
Militaire [...]	304	0.92	0.01	0.92	Médailles	28	0.84	0.00	0.79
Agriculture [...]	259	0.80	0.04	0.80	Caractères	27	0.67	0.00	0.51
Belles-lettres - Poésie	246	0.75	0.01	0.74	Politique	27	0.31	0.00	0.00
Anatomie	245	0.92	0.02	0.91	Minéralogie	26	0.68	0.00	0.65
Mathématiques	164	0.88	0.00	0.89	Superstition	26	0.81	0.00	0.73
Musique	163	0.94	0.01	0.94	Spectacle	11	0.17	0.00	0.00

TAB. 3: F-mesures obtenues par ensemble de domaines avec les approches SGD + TF-IDF (1), LSTM + FastText (2) et BERT (3) sans échantillonnage et sur le jeu de test.

Cette disparité entre les classes est visible dans le tableau 3 qui présente les scores (f-mesure) obtenus sur le jeu de test (sans échantillonnage) pour chaque classe (triées selon leur nombre d'articles) avec les méthodes SGD + TF-IDF (1), LSTM + FastText (2) et BERT (3). Sur les 38 classes, 30 obtiennent plus de 70% avec BERT (31 avec SGD+TF-IDF et aucune avec LSTM+FastText) et 2 ont moins de 50% avec BERT (3 avec SGD+TF-IDF et toutes avec LSTM+FastText). On remarque de manière générale que les classes les plus peuplées (> 1 000 articles) obtiennent de très bons scores, telle que la classe *Géographie* avec 99%. Pour les classes les moins peuplées, on constate une baisse des scores. On note néanmoins quelques exceptions, comme par exemple les classes *Chasse* et *Pêche* qui ont peu d'articles (124 et 48) mais qui sont bien classées (92% et 93% avec BERT). On constate que les mauvais résultats obtenus par l'entraînement d'un réseau LSTM sont confirmés et même accentués sur cette répartition par classe où seulement 2 classes obtiennent plus de 1% (respectivement 22% et 39%). Au delà du nombre d'articles par classe et donc du manque de données pour certaines d'entre elles (voir même pour la totalité dans le cadre des approches par apprentissage profond), ces chiffres mettent en avant la difficulté de catégoriser ou distinguer certaines classes, en particulier à cause de leur proximité lexicale ou sémantique.

La figure 3 présente la matrice de confusion obtenue avec la méthode SGD+TF-IDF sur le jeu de test. On peut voir qu'un grand nombre d'articles des classes *Arts et métiers* et *Economie*

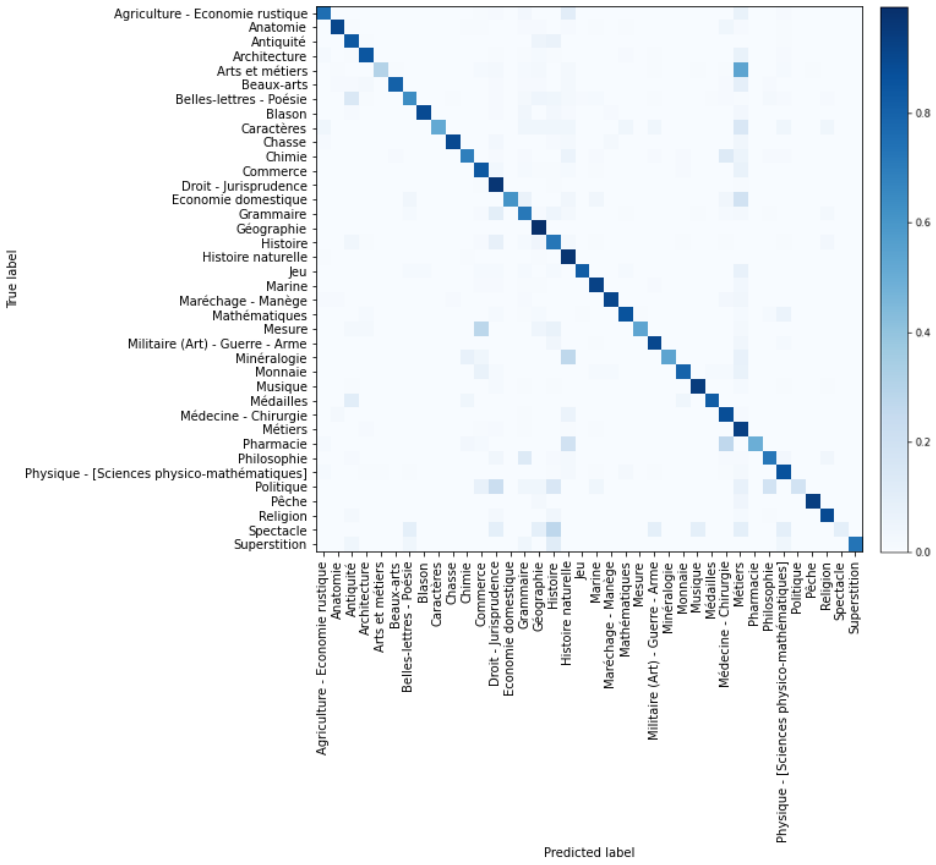

FIG. 3: Matrice de confusion obtenue avec l'approche SGD+TF-IDF sur le jeu de test.

domestique a été classé dans la classe *Métiers*, de la même manière les classes *Mesure*, *Minéralogie*, *Pharmacie* et *Politique* sont souvent confondues avec les classes *Commerce*, *Histoire naturelle*, *Médecine - Chirurgie* et *Droit - Jurisprudence*, respectivement. Les proximités sémantiques entre ces classes montrent bien la difficulté pour les modèles de choisir entre l'une ou l'autre et les résultats confirment qu'en cas de trop grande proximité les modèles choisissent la classe la plus représentée dans le jeu de données.

4 Conclusion

Dans cet article nous avons présenté une étude comparative de différentes approches de classification supervisée appliquées à la classification automatique d'articles encyclopédiques. À partir du corpus de l'Encyclopédie de Diderot et d'Alembert et de la classification en ensemble de domaines proposée par l'ENCCRE, nous avons pu entraîner différents modèles de classification. Nous avons expérimenté différentes approches de vectorisation des textes (sac

de mots, TF-IDF, doc2vec, FastText) combinées à des méthodes d'apprentissage automatique classiques (Naive Bayes, Logistic regression, Random forest, SGD et SVM) ou d'apprentissage profond (CNN, RNN et LSTM). Enfin, nous avons également spécialisé des modèles de langue pré-entrainés utilisant une architecture à base de *transformers* (BERT et CamemBERT) pour notre tâche de classification. Nous avons évalué nos différentes expérimentations sur des jeux de validation et de test pour différentes configurations (échantillonnage du nombre d'articles pour équilibrer les classes, et choix des hyperparamètres des algorithmes de classification). En terme de classification, les résultats sont très encourageants et un grand nombre de classes sont bien identifiées par les modèles (22 / 38 obtiennent plus de 70% de f-mesure). Les modèles de langue dynamiques pré-entrainés tels que BERT et CamemBERT obtiennent des résultats similaires aux méthodes SGD et Logistic Regression associées à une vectorisation TF-IDF (89% de f-mesure moyenne). Par ailleurs, nos résultats viennent confirmer les différents travaux de la littérature selon lesquels les modèles BERT permettent d'obtenir de bonnes performances avec moins de données. En effet, dans notre cas, les méthodes d'apprentissage profond CNN et LSTM obtiennent de très mauvais résultats à l'inverse des méthodes d'apprentissage classiques ou de BERT. Enfin, les résultats ont montré la difficulté de distinguer certaines classes, en partie du fait de leur proximité sémantique ou lexicale, mais également à cause de l'important déséquilibre entre elles. Les modèles de type BERT obtiennent des résultats plus élevés mais certaines classes restent difficiles à identifier et sont confondues avec des classes proches sémantiquement et plus représentées dans le jeu de données. De plus, les mauvais résultats obtenus pour certaines classes sont à modérer. En effet, les choix de classification fait par les auteurs ou les contributeurs de l'ENCCRE peuvent être remis en question. Les différences de classification obtenues avec les modèles peuvent alors servir de point d'entrée pour une étude qualitative. C'est alors au littéraire, au linguiste, ou à l'historien des idées de prendre le relais pour s'interroger sur les motifs qui ont menés au choix de l'auteur.

Notre objectif principal est maintenant d'appliquer notre meilleur modèle pour la classification d'articles issus de différentes encyclopédies. Certaines, telles que La Grande Encyclopédie (1886-1902) n'ont pas d'indication de domaine au niveau des articles. Dans un contexte d'analyse, de recherche et d'extraction d'information, il est important de pouvoir identifier la thématique des articles en amont des différents traitements. Enfin, certaines autres encyclopédies proposent une classification. Dans ce cas, notre objectif est de la comparer avec les résultats de notre classification automatique afin d'étudier les changements ou les divergences. Cela nous permettra dans certains cas de corriger ou enrichir des classifications établies (correction d'erreurs de classification, ou ajout dans le cas d'un article appartenant à plusieurs domaines). Les modèles de classifications obtenus nous permettrons également de faire une étude diachronique des changements survenus au cours des cours des siècles et ainsi analyser l'évolution des discours encyclopédiques.

Remerciement

Les auteurs remercient le LABEX ASLAN (ANR-10-LABX-0081) de l'Université de Lyon pour son soutien financier dans le cadre du programme français "Investissements d'Avenir" géré par l'Agence Nationale de la Recherche (ANR).

Références

Bojanowski, P., E. Grave, A. Joulin, et T. Mikolov (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics 5*, 135–146.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

Guilbaud, A. (2017). L'enccre, édition numérique collaborative et critique de l'encyclopédie. *Recherches sur Diderot et sur l'Encyclopédie* (52), 5–22.

Horton, R., R. Morrissey, M. Olsen, G. Roe, R. Voyer, et al. (2009). Mining eighteenth century ontologies : machine learning and knowledge classification in the encyclopédie. *3*(2).

Le, Q. et T. Mikolov (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196. PMLR.

Martin, L., B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, et B. Sagot (2019). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*.

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

Roe, G., C. Gladstone, et R. Morrissey (2016). Discourses and disciplines in the enlightenment : Topic modeling the french encyclopédie. *Frontiers in Digital Humanities 2*, 8.

Salton, G. et M. J. McGill (1986). *Introduction to Modern Information Retrieval*. USA : McGraw-Hill, Inc.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, et I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Summary

This article proposes a comparative study of different supervised classification approaches applied to the automatic classification of encyclopaedic articles. Our training corpus is composed of the 17 volumes of text of the Encyclopédie by Diderot and d'Alembert (1751-1772) representing a total of about 70,000 articles. We have experimented different approaches for text vectorization (bag of words and word embeddings) combined with classical machine learning methods, deep learning and BERT architectures. In addition to the comparison of these different approaches, our objective is to automatically identify the domains of the unclassified articles of the Encyclopaedia (about 2400 articles). The best model obtains 89% of average f1-score for the 38 classes. Moreover, our study highlights the difficulty of classifying certain semantically close classes. All the code developed and the results obtained in the framework of this project are available in open-source.