Attribution d'auteur par utilisation des méthodes d'apprentissage profond

Trang Lam, Jérémy Demange, Julien Longhi

Attribution d'auteur par utilisation des méthodes d'apprentissage profond

Trang Lam*, Jérémy Demange**
Julien Longhi***

*Trang Lam, CY Cergy Paris Université,
Laboratoire IDHN, AGORA,
33, boulevard du Port 95011 Cergy-Pontoise,
doantrang1110@gmail.com
**Jérémy Demange, CY Cergy Paris Université,
Laboratoire IDHN,
33, boulevard du Port 95011 Cergy-Pontoise,
jeremy.demange@cyu.fr
***Julien Longhi, CY Cergy Paris Université,
Laboratoire IDHN, AGORA, Institut Universitaire de France,
33, boulevard du Port 95011 Cergy-Pontoise,
julien.longhi@cyu.fr

Résumé. L'attribution d'auteur est un sous-domaine du Traitement Automatique des Langues (TAL) qui consiste à identifier l'auteur le plus probable d'un texte parmi un ensemble de candidats. Ainsi, elle peut être vue comme un problème de classification de textes. Cette tâche fait partie des sciences forensiques, de la détection du plagiat, voire l'identification des cybercriminels. Il s'agit d'un domaine interdisciplinaire avec un croisement de l'apprentissage automatique et de la stylométrie. La plupart des travaux précédents se focalisent sur des textes longs tandis que les tendances actuelles des technologies de l'information encouragent le recours à des textes plus courts et informels. Ainsi, cette contribution vise à étudier la question d'attribution d'auteur sur des messages courts extraits du service micro-blogging Twitter. Pour cette étude, nous nous servons des CNN [1] et des LSTM [2].

1 Introduction

A ce jour, il n'est pas pratiqué d'analyse linguistique sur les textes soumis aux techniciens travaillant pour les tribunaux français. L'attribution d'auteur par analyse linguistique constituerait donc un nouveau champ d'analyse dans le domaine des Documents et pourrait à terme aboutir à la mise en place d'une nouvelle spécialité criminalistique. Pour ce faire, elle

1. Réseaux de neurones convolutionnels.
2. Réseaux récurrents à mémoire court et long terme.

doit démontrer sa capacité à répondre aux problématiques propres à l'environnement dans lequel la criminalistique évolue, afin d'être utilisable. Les spécialistes en analyse de documents dans le domaine de la criminalistique analysent des traces liées aux documents ou autres objets assimilés. Ces traces proviennent, par exemple, des titres sécurisés (passeports, permis de conduire...), des lettres anonymes (manuscrites, tapuscrites ou dactylographiées), des contrats, des chèques, des tags, des testaments ou encore des formulaires administratifs. Les examens peuvent porter sur la composition du papier, la recherche de traces mécaniques non visibles à l'œil nu (en général des mentions manuscrites ou des marques des galets d'entraînement d'une imprimante), la détermination des techniques d'impression, l'analyse et la différenciation d'encres et la comparaison d'écritures manuscrites. La délimitation du terrain de travail pour le linguiste constitue une porte d'entrée essentielle pour appréhender les possibles champs d'action d'une discipline encore en construction, et répondre aux enjeux des criminalisticiens - Longhi (2021). Le but de cet article est de rendre compte des avancées obtenues lors de la réalisation d'un projet de recherche (CHEMI IRITA) mené à propos de l'analyse textuelle d'écrits et l'attribution d'auteurs, afin d'apporter un nouvel axe de recherche concernant des textes analysés dans le domaine de la comparaison d'écritures en criminalistique. Avec les évolutions du TAL, notamment par le biais de l'apprentissage profond grâce aux réseaux de neurones, des nouvelles perspectives sont à envisager pour avancer dans ce domaine. L'attribution d'auteur est un sous-domaine du TAL qui consiste à identifier l'auteur le plus probable d'un texte parmi un ensemble de candidats. Ainsi, elle peut être vue comme un problème de classification de textes, dont l'apprentissage supervisé. Il s'agit d'un croisement de la stylométrie et de l'apprentissage automatique.

La stylométrie est une branche de la linguistique computationnelle qui étudie le style littéraire à l'aide des méthodes quantitatives. Elle suppose qu'un auteur laisse de façon inconsciente dans son texte des caractéristiques qui peuvent mener à son identification. Les caractéristiques stylométriques utilisées en Attribution d'auteur peuvent être séparées en différents niveaux d'analyse (lexical, syntaxique, sémantique, structurel et spécifique aux applications). Ces caractéristiques peuvent être : le nombre de mots, de phrases ; la longueur de mots, de phrases ; la fréquence des mots outils, des mots-formes, des n-grammes de mots, des n-grammes de caractères ; la fréquence des parties du discours ; les collocations (la fréquence des bigrammes de parties de discours) ; le nombre de hapax legomena (mots apparaissant une seule fois) ; etc. Un état de l'art complet des caractéristiques stylométriques peut être consulté dans les travaux de Stamatatos (2009), de Bouanani et Kassou (2014) et de Lagutina et al. (2019). En réalité, il n'y a pas de consensus pour dire quelles sont les caractéristiques les plus optimales et la performance de la tâche d'attribution d'auteur dépend considérablement des caractéristiques stylométriques choisies. Ainsi, il est nécessaire d'avoir recours à des techniques fiables et efficaces pour extraire des caractéristiques appropriées. Pour notre cas d'étude, comme nous travaillons sur des tweets, des n-grammes de mots et des n-grammes de caractères apparaissent comme étant les caractéristiques les plus correspondantes vu qu'elles sont moins liées au type et au sujet de textes. En plus, ils sont capables de capturer les informations lexicales, contextuelles ou thématiques (pour des grandes valeurs de n) sans avoir besoin des connaissances préalables de la grammaire d'une langue. En d'autres termes, ils peuvent s'appliquer aux différentes langues naturelles - Stamatatos (2006) ; Grieve (2007). Stamatatos (2009) montre aussi que des n-grammes de caractères ne sont pas trop affectés par des textes « bruités ». Grâce à sa robustesse, cette approche se trouve dans plusieurs travaux et donne des résultats positifs -

Stamatatos (2009, 2013) ; Layton et al. (2010) ; Tanguy et al. (2011) ; Sun et al. (2012)..

A propos des méthodes d'apprentissage, différentes techniques pour exploiter les caractéristiques stylométriques ont été proposées : SVM (Machine à vecteurs de support), Decision Trees (Arbres de décision), Random Forest (Forêts aléatoires), Naive Bayes et les réseaux de neurones. Parmi elles, les méthodes utilisant les réseaux de neurones connaissent un véritable essor au cours des dernières années, dépassent dans certains cas les méthodes statistiques traditionnelles et donnent des résultats spectaculaires.

2 Données et méthodes

2.1 Le corpus

Les données utilisées dans notre protocole proviennent du corpus des tweets de la #présidentielle2017 [3]. Il est vrai que les tweets politiques ne sont pas nécessairement représentatifs des tweets du grand public. Néanmoins, nous constatons que ce type de tweets nous fournit quand même les caractéristiques appropriées pour que nous puissions voir la potentialité de ce sujet de recherche - Longhi (2017). Nous avons choisi ces données car il s'agit d'un corpus exploratoire bien connu de divers travaux qui permet de tester de manière fiable les réseaux. Le corpus #presidentielle 2017 comprend au total 42923 tweets de 11 candidats produits lors des élections de 2017 (voir la figure 1). Ce corpus est divisé ensuite en 3 sous-corpus avec un ratio 60-20-20 qui permettent d'entraîner, estimer et évaluer le modèle.

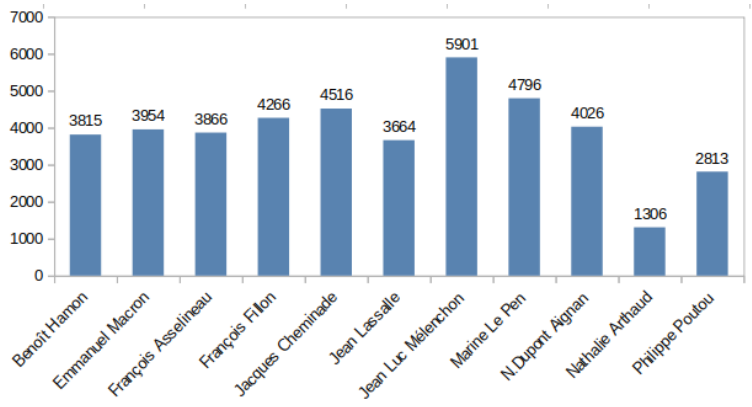

FIG. 1 – *Distribution des tweets par candidat.*

2.2 Pré-traitements

Avant d'être modélisées, les données du corpus passent par une étape de prétraitement qui joue un rôle considérable à la performance du modèle. Cette étape consiste en nettoyage des données : la suppression des liens, des traces des balises HTML, des caractères de ponctuation ;

3. https ://www.ortolang.fr/market/corpora/corpus-presidentielle2017

la séparation de la symbole # dans les hashtags. Ainsi, les hashtags sont considérés comme des mots simples.

2.3 Les types des réseaux de neurones et comparaisons

Au cours de nos expérimentations, nous avons construit deux modèles de réseaux de neurones les plus connus pour la tâche de classification en général et en particulier celle d'attribution d'auteur : les réseaux de neurones convolutionnels (CNN) et les réseaux à mémoire court et long terme (LSTM).

2.3.1 Les réseaux de neurones convolutionnels (CNN)

Notamment connus comme les plus performants modèles pour le traitement des images, ces réseaux ne connaissent un essor dans le traitement du langage naturel qu'à partir des travaux de Kim (2014), Kalchbrenner et al. (2014), Zhang et al. (2017), Conneau et al. (2017) pour la tâche de classification des textes. En TAL, ils sont considérés comme une généralisation des modèles n-grammes. Plus concrètement, les CNNs utilisent un mécanisme de filtres qui a pour but de détecter des motifs récurrents en faisant glisser des fenêtres sur l'ensemble du texte. Avec k filtres de tailles égales ou différentes, plusieurs types de motifs particuliers peuvent être capturés. En plus, le back-propagation permet d'accorder à des points à des éléments linguistiques qui contribuent au mécanisme décisionnel du modèle. Deux architectures CNNs sont proposées pour notre protocole d'expérimentation, notées respectivement CNN-1, CNN-2.

— CNN-1 constitue une architecture dite "traditionnelle" de CNN. Elle est composée d'une couche de pré-apprentissage [4] (ou Word Embbeding), d'une couche de convolution suivie d'une couche de max-pooling. La sortie de ce bloc convolutif passe ensuite à une couche entièrement connectée, une couche DropOut et une dernière couche entièrement connectée associée à la fonction d'activation Softmax.

— CNN-2 s'inspire des travaux de Kim (2014). Au lieu d'utiliser une seule couche de convolution, nous utilisons 3 couches de convolutions mises en parallèle avec des filtres de taille 3,4 et 5 (100 filtres pour chaque taille). Le reste du réseau reste inchangé en comparaison avec CNN-1.

La qualité de l'apprentissage peut être affectée par les hyperparamètres qui sont des paramètres dont la valeur doit être définie avant le processus d'entraînement du modèle. Ces hyperparamètres sont relatifs soit à l'architecture du modèle (nombre de couches, nombre de neurones par couche, etc.) soit à l'entraînement du modèle (batch size, taux d'apprentissage, époque, etc.). Afin de trouver les valeurs optimales des hyperparamètres, nous avons dû effectuer plusieurs essais. Le tableau 1 définit les valeurs optimales de deux modèles CNN après bon nombre de tests.

2.3.2 Les réseaux à mémoire court et long terme (LSTM)

Les LSTMs sont un des types de réseaux de neurones récurrents. Ces réseaux sont des modèles spécialisés dans le traitement des données de type séquentiel ou temporel. La particularité

4. Dans notre cas d'étude, nous avons recours à des Word Embeddings pré-entraînés de FastText avec la taille de vocabulaire fixée à 30000 et la taille d'Embedding fixée à 300.

	CNN-1	CNN-2
Nombre de filtres	300	100 filtres pour chaque taille
Taille de filtres	3	2,3,4
L2 (valeur lambda)	0,00001	None
Nombre de neurones	128	128
Dropout	0,5	0,5
Dropout	0,5	0,5
Taux d'apprentissage	0,0005	0,001
Nombre d'époques	20	20
Arrêt prématuré	3	3
Taille de batch	2000	2000

TAB. 1 – *La configuration optimale des CNNs*

de ce type d'architecture réside dans sa capacité à modéliser les dépendances entre les mots. A l'aide de leur mémoire interne, ces réseaux sont capables d'utiliser l'information contextuelle passée lors du traitement de l'information courante dans une séquence. D'une manière succincte, ils vont parcourir une séquence textuelle, par exemple de gauche à droite, un mot après l'autre et mettre à jour leur mémoire interne à chaque pas. Ce type de réseaux est performant pour diverses applications de TAL telles que : la modélisation du langage ; la reconnaissance des entités nommées ; l'étiquetage morpho-syntaxique ; la prédiction du prochain caractère, mot d'une séquence textuelle ; l'analyse du sentiment ; la classification de textes, etc. Nous construisons en total 4 modèles LSTM, dont 2 modèles du LSTM unidirectionnel (LSTM-1 et LTSM-2) et 2 modèles du LSTM bidirectionnel (BiLSTM-1 et BiLSTM-2).

LSTM unidirectionnel prend en compte seulement le contexte précédent lors du traitement de l'élément courant dans une séquence.

— LSTM-1 est composé d'une couche de pré-apprentissage, d'une couche de LSTM, d'une couche entièrement connectée, d'une couche Dropout et d'une couche entièrement connectée associée à la fonction d'activation Softmax.
— LSTM-2 constitue une variante de LSTM-1. Au lieu de prendre seulement le vecteur du dernier état caché (last hidden state), nous prenons tous les vecteurs des états cachés et les faisons entrer dans la couche Max Pooling.

LSTM bidirectionnel est capable de prendre en compte à la fois le contexte passé et le contexte futur.

— BiLSTM-1 est composé d'une couche de LSTM forward, d'une couche de LSTM backward et enfin des couches entièrement connectées pour la classification.
— BiLSTM-2 est une variante de BiLSTM-1. Ce modèle retourne tous les vecteurs intermédiaires et les fait passer à la couche Max Pooling.

Nous avons testé les différentes valeurs pour différents hyperparamètres. Et la table 2 est un récapitulatif des valeurs optimales du modèle.

	LSTM	BiLSTM
Couche LSTM forward	75	100
Couche LSTM backward	75	100
Nombre de neurones	96	64
Dropout	0,5	0,5
Taux d'apprentissage	0,01	0,01
Nombre d'époques	20	20
Arrêt prématuré	3	3
Taille de batch	2000	2000

TAB. 2 – *La configuration optimale des LSTMs*

Une fois que les modèles sont construits, nous les appliquons au jeu de test qui contient 8456 tweets. Le tableau 3 et la figure 2 nous montrent les résultats de six modèles en termes de taux d'exactitude, précision, rappel et f-mesure. Parmi les modèles proposés, le modèle CNN-2, le modèle CNN avec 3 couches de convolution mises en parallèle, est le plus performant avec un taux d'exactitude de 83%. Néanmoins, les résultats obtenus par le modèle CNN-1 sont aussi très remarquables. Étant un simple CNN constitué d'une seule couche de convolution, nous observons tout de même des bons résultats avec un taux d'exactitude de 81%. Quant aux réseaux LSTMs, les résultats obtenus ne sont pas très satisfaisants. Le taux d'exactitude de LSTM-1, LSTM-2, BiLSTM-1, BiLSTM-2 sont, respectivement de 73%, 78%, 77% et 81%. Pourtant, à partir de ces résultats, nous pouvons faire quelques remarques :

— Les LTSMs bidirectionnels sont plus performants que les LTSMs unidirecionnels.
— L'approche d'utiliser tous les états cachés (LSTM-2 et BiLSTM-2) donne les meilleurs résultats que celle d'utiliser seulement le dernier état caché (LSTM-1 et BiLSTM-1).

	Exactitude	Précision	Rappel	F-Mesure
CNN-1	0.81	0.82	0.80	0.81
CNN-2	0.83	0.82	0.82	0.83
LSTM-1	0.73	0.73	0.72	0.72
LSTM-2	0.78	0.77	0.77	0.77
BiLSTM-1	0.77	0.76	0.76	0.76
BiLSTM-2	0.81	0.80	0.79	0.80

TAB. 3 – *Les résultats de six modèles en attribution d'auteur des tweets en termes de taux d'Exactitude, Précision, Rappel et F-mesure*

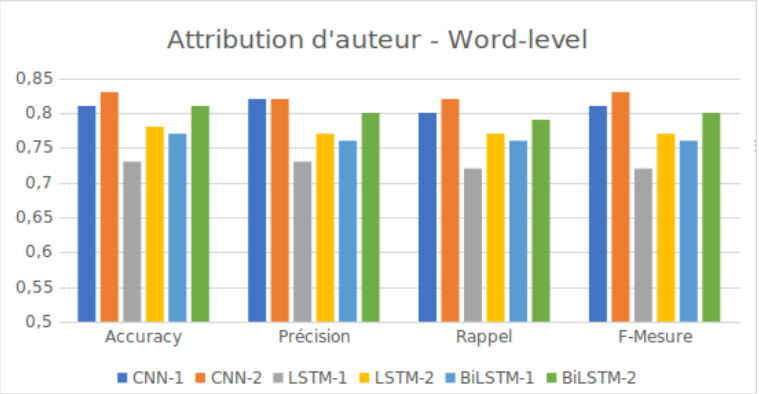

FIG. 2 – *Comparaison de six modèles pour l'attribution d'auteur des tweets*

2.4 Comparaison avec d'autres algorithmes d'apprentissage

Dans l'idée de savoir si les techniques d'apprentissage profond peuvent dépasser les techniques traditionnelles, nous avons implémenté quelques méthodes d'apprentissage traditionnelles les plus utilisées pour la classification telles que : Naive Bayes, SVM, Forêts Aléatoires et Arbre à décision.

	Exactitude
Naive Bayes	0.68
SVM	0.76
Random Forest	0.74
Decision Tree	0.68

TAB. 4 – *Les résultats des algorithmes d'apprentissage traditionnels*

Comme le montre le tableau 4, parmi les quatre algorithmes d'apprentissage traditionnels testés, SVM obtient le meilleur résultat avec un taux d'exactitude de 76%. On peut donc en déduire que les modèles d'apprentissage profond peuvent surpasser la performance des techniques traditionnelles dans la tâche d'attribution d'auteur. Pour un travail ultérieur, nous envisageons de combiner l'usage de Word Embeddings avec les techniques d'apprentissage traditionnelles afin d'avoir une comparaison plus fiable entre les approches.

3 Résultats

Parmi les six modèles proposés, nous avons choisi le modèle CNN-2 pour faire une analyse plus détaillée. Pour avoir une idée générale sur la classification des données du corpus de test, nous les avons projeté dans une espace de deux dimensions à l'aide de t-SNE (t-Distributed

Stochastic Neighbor Embedding). Il s'agit d'une technique de réduction de dimension qui permet de projeter des données à haute dimension dans un espace de deux ou trois dimensions.

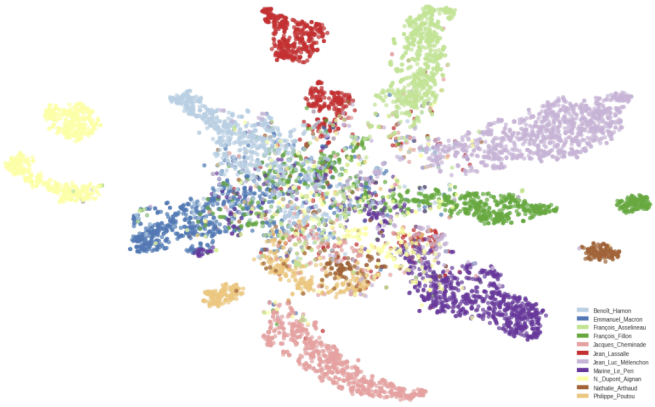

FIG. 3 – *Visualisation du regroupement des tweets du corpus de test*

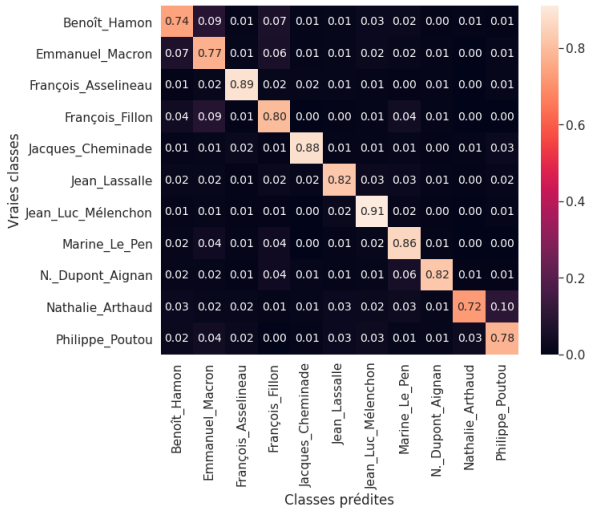

FIG. 4 – *Matrice de confusion du modèle CNN-2*

Comme le montrent la figure 3 et la figure 4, les tweets de Jacques Cheminade, Jean-Luc Mélenchon, François Asselineau, Nicolas Dupont-Aignan et Jean Lassalle sont bien regroupés. Dans le cas de Nathalie Arthaud, certains de ses tweets sont reconnus par l'algorithme comme proches des tweets de Phillipe Poutou. De même, un échantillon des tweets de Nicolas Dupont-Aignan se trouve proche des tweets de Marine Le Pen. En plus, nous remarquons que le modèle

rencontre des difficultés dans la distinction des tweets de Benoît Hamon, Emmanuel Macron et François Fillon. La confusion entre les tweets de Marine Le Pen et ceux de François Fillon est aussi remarquable. Des exemples de tweets qui ne sont pas correctement classifiés seront détaillés dans la partie suivante.

4 Discussion

4.1 Amélioration des pré-traitements

Il est toujours possible d'améliorer l'étape du pré-traitement. Dans l'algorithme actuel, nous avons supprimé certaines données (comme des liens, des caractères de ponctuation). Cependant, Twitter permet d'écrire des messages assez courts, limités à 280 caractères aujourd'hui. Nous pouvons donc nous poser la question de savoir s'il est légitime de laisser sans pré-traitements une grande majorité des données et essayer de ne pas supprimer la ponctuation ou d'autres caractéristiques par exemple. Avec si peu de caractères, il est peut-être préférable de garder un maximum de caractéristiques plutôt que d'en retirer.

4.2 Explication sur les résultats

Dans l'idée de comprendre le mécanisme décisionnel du modèle, nous nous basons sur la mesure de similarité. Plus concrètement, nous supposons que quand le modèle attribue à un tweet un auteur A, cela signifie que l'algorithme reconnaît une similitude entre ce tweet et un tweet de cet auteur dans le jeu d'entraînement. Dans cette idée, nous avons utilisé la similarité cosinus pour trouver, dans la base de données d'entraînement, le tweet le plus similaire du tweet en question. Prenons par exemple un tweet de Marine Le Pen : « La fermeture de #Fessenheim est une décision idéologique faite sous la pression des écologistes. ». Ce tweet est attribué par l'algorithme à Emmanuel Macron. En ayant recours au mesure cosinus, nous trouvons que ce tweet est considéré "proche" au tweet : « La fermeture de Fessenheim est une décision responsable que je maintiendrai », un tweet d'Emmanuel Macron qui se trouve dans le jeu d'entraînement. Un autre exemple est un tweet de François Fillon : « Je veux que, d'ici la fin du quinquennat, 100% du territoire soit couvert en Très Haut Débit fixe et mobile. ». L'algorithme attribue ce tweet à Emmanuel Macron en référence à son tweet : « Ce projet, c'est le vôtre : la couverture du territoire en très haut débit sera une des priorités de ce quinquennat ». Cela montre à la fois la complexité de la tâche (car il y a des similarités stylistiques, thématiques entre auteurs) et l'intérêt d'analyser et de comprendre les résultats, qui enrichissent la connaissance du thème de recherche.

4.3 Intégration dans une application Web

Nous avons pu intégrer cet algorithme dans une application Web (voir figure 5). Il s'agit de TextPrint, une application que nous avons développé afin de stocker et gérer des corpus de texte, de faire des analyses de celles-ci. L'application contient également des outils et algorithmes pour de la comparaison d'auteurs. Nous ne pouvons pas détailler davantage l'application ici, car il s'agit de travaux en cours qui visent à être utilisé dans une variété d'applications et seront également amenés à être améliorés.

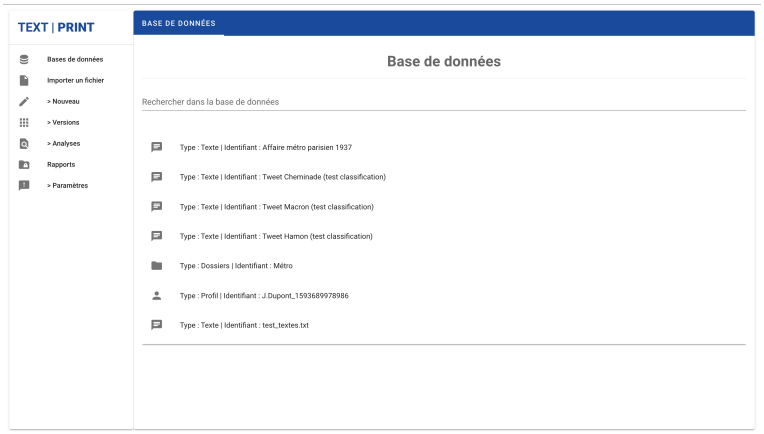

FIG. 5 – *Application TextPrint*

4.4 Choix du corpus

Nous avons choisi pour cette recherche un corpus de tweets politiques (en contexte électoral). Ce type de communication bénéficie d'une certaine unité temporelle (élection 2017), thématique (en fonction de la campagne) et a par ailleurs été décrit dans d'autres travaux, ce qui aide à analyser les résultats obtenus. Bien sûr, la question de la prise en charge des messages par un-e ou plusieurs communicant-e-s peut se poser, de même que des tentatives de démarcation à tel ou tel stade de la campagne par tel ou tel candidat. Pour enrichir notre recherche, nous travaillons actuellement sur la récolte d'articles de presse écrits par des auteurs clairement définis (comme le genre de l'éditorial). Ceci nous permettra d'affiner nos analyses et d'améliorer en conséquence notre modèle.

5 Conclusions

Dans cet article, nous avons proposé différents modèles de réseaux de neurones pour la tâche d'attribution d'auteur, dont 2 modèles CNN et 4 modèles LSTM. Nous avons pu montrer que l'application des méthodes d'apprentissage profond à cette tâche est une direction prometteuse avec un taux d'exactitude à 83% (modèle CNN-2). Ces résultats peuvent sans doute être améliorés en modifiant l'algorithme de pré-traitement comme évoqué dans la partie discussion. Il est à noter que nous avons utilisé comme données d'entraînement des tweets politiques, mais il est possible de l'appliquer à d'autres types de tweets ou textes courts afin que l'on soit plus proches du terrain d'analyse souhaité. Face à la croissance exponentielle des textes anonymes sur Internet, plus spécifiquement sur des réseaux sociaux, le problème d'authentification d'auteurs devient plus urgent que jamais.

Le travail que nous avons présenté ici n'est qu'un premier pas vers le domaine d'attribution d'auteur associé à l'apprentissage profond. Comme perspectives de ce travail, plusieurs pistes peuvent être envisagés :

— Au lieu d'utiliser un seul trait linguistique, nous allons combiner plusieurs traits linguistiques en même temps (par exemple : mots + lemmes + parties du discours) avec l'espoir que cela va rendre le système d'attribution d'auteur plus robuste. Il serait par exemple possible de tenir compte des apports de méthodes textométriques comme dans Longhi (2021), en les croisant avec l'approche Deep Learning présentée ici, comme le proposent pour d'autres applications Vanni et al. (2018).
— Segmenter les tweets non pas au niveau de mots mais au niveau de lettres pour garder les majuscules, les ponctuations parce que ces éléments sont aussi utiles pour quantifier le style d'un auteur.
— Focaliser sur l'étape de prétraitement des données, dont la normalisation. Comme notre corpus est constitué de tweets politiques, il ne contient pas beaucoup des éléments «bruyants». Pourtant, les tweets en général se caractérisent par les abréviations, les émoticônes, les mots allongés, etc.
— Construire des modèles plus «profonds» par exemple : combiner les réseaux de neurones convolutionnels (CNN) avec les réseaux récurrents à mémoire court et long terme (LSTM) ; empiler les couches LSTM et ajouter une couche d'Attention. En plus, nous voudrions tester le modèle BERT qui définit l'état de l'art pour plusieurs tâches de TAL.
— Augmenter le volume de corpus ainsi que le nombre d'auteurs. Une fois que nous obtiendrons un corpus assez grand, nous pourrons aussi entraîner notre propre Word Embeddings au lieu d'utiliser les Word Embeddings pré-entraînés.

6 Remerciements

Le projet IRITA (Inventaires des Ressources Informatiques et Textométriques pour l'Attribution d'auteurs) était porté par Julien Longhi, et était composé de Jérémy Demange, Alexandra Freeman, et Trang Lam au sein de l'Institut des humanités numériques de CY Cergy Paris Université. La région Île-De-France, par l'appel du DIM Sciences du textes et connaissances nouvelles, soutient également ce projet (PhD2). Financé dans le cadre de l'AAP CHEMI (Centre des Hautes Études du Ministère de l'Intérieur, thème II de l'appel : « L'impact de la révolution numérique sur les métiers de la sécurité »).

Références

Bouanani, S. E. M. E. et I. Kassou (2014). Authorship analysis studies: A survey. *International Journal of Computer Applications 86*, 22–29.

Conneau, A., H. Schwenk, L. Barrault, et Y. Lecun (2017). Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain, pp. 1107–1116. Association for Computational Linguistics.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing 22(3)*, 251–270.

Kalchbrenner, N., E. Grefenstette, et P. Blunsom (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, pp. 655–665. Association for Computational Linguistics.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1746–1751. Association for Computational Linguistics.

Lagutina, K., N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, I. Paramonov, et P. G. Demidov (2019). A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pp. 184–195.

Layton, R., P. Watters, et R. Dazeley (2010). Authorship attribution for twitter in 140 characters or less. In *2010 Second Cybercrime and Trustworthy Computing Workshop*, pp. 1–8.

Longhi, J. (2017). Humanités, numérique : des corpus au sens, du sens aux corpus. *Questions de communication*.

Longhi, J. (2021). Using digital humanities and linguistics to help with terrorism investigations. *Forensic Science International*, 110564.

Stamatatos, E. (2006). Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools 15(5)*, 823–838.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal American Society for Information Science and Technology 60(3)*, 538–556.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law & Policy 21*, 421–725.

Sun, J., Z. Yang, S. Liu, et P. Wang (2012). Applying stylometric analysis techniques to counter anonymity in cyberspace. *Journal of Networks 7*.

Tanguy, L., A. Uriely, B. Calderone, N. Hathout, et F. Sajous (2011). A multitude of linguistically-rich features for authorship attribution. In *Notebook for PAN at CLEF 2011*, Amsterdam, Netherlands.

Vanni, L., D. Mayaffre, et D. Longrée (2018). ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables. In *JADT 2018*, Rome, Italy.

Zhang, T., C. Li, N. Cao, R. Ma, S. Zhang, et N. Ma (2017). Text feature extraction and classification based on convolutional neural network (cnn). pp. 472–485.

Summary

Authorship attribution is a subfield of Natural Language Processing (NLP) that identifies the most possible author of a text among a group of candidate authors. Thus, it can be seen as a text classification problem. This task is part of forensic sciences, detecting plagiarism or cybercriminal analysis. This is an interdisciplinary research that involves machine learning and stylometry. Most of the previous studies focus on long texts while current trends in information technology encourage shorter and informal texts. Therefore, this contribution aims to tackle the task of identifying authors of tweets, short messages from the most common micro-blogging service. For this study, two types of neural networks will be used: CNN and LSTM.