The 38th Canadian Conference on Artificial Intelligence
DOI: 00.000
Canadian AI 2025
26—29 May 2025, Calgary, Alberta

Démarche de développement d'un algorithme prédictif des profils OCEAN d'utilisateurs de Reddit

Oumar Bocoum†, Jordan Carré†, Bineta Diallo †, Antoine Jean †
† Electrical and Computer Engineering, Université Laval, Canada, QC

Abstract

Cette étude documente le processus de production d'un système de prédiction des scores au test de personnalité OCEAN (Big Five) d'utilisateurs de Reddit. Elle utilise un *pipeline* combinant l'analyse sémantique des messages, le traitement des métadonnées et des modèles d'apprentissage automatique (forêts aléatoires et analyse de sentiment). Cette approche s'inscrit dans la lignée des travaux de Jean et al. [1] sur l'analyse de données massives appliquée à la psychométrie. Bien que les résultats présentent une précision globale satisfaisante, des variations importantes de précision persistent entre différentes manières de prédire les scores OCEAN, suggérant que les commentaires de la personnes sont moins révélateurs de sa personnalité que les informations contenues dans son profil.

Keywords: Big Data, Natural Language Processing (NLP), OCEAN Personality Prediction, Pushshift dataset

Introduction

Reddit est un site web d'actualités et de discussions structuré autour de communautés thématiques «subreddits» où les utilisateurs sont invités à échanger sur des intérêts communs. Avec environ 2 milliards de visites mensuelles, la plateforme représente une source précieuse d'information sur le comportement et la psychologie des internautes. Au moyen de l'API Pushshift, Baumgartner et al.[2] ont créé un jeu de données compilant l'ensemble des messages postés sur le site de 2015 à 2019. Travaillant à partir des profils des utilisateurs réalisés par Gjurković et al. [3] sur ce corpus, nous avons élaboré une méthodologie visant à prédire ces scores en se basant sur les données contenues dans les deux corpus [1]. Nous détaillons ici la suite du processus d'implémentation de l'architecture, allant de l'implémentation des algorithmes jusqu'à la prédiction finale des scores.

Sélection des algorithmes

Analyse des sentiments

Puisque plusieurs échelles du score OCEAN sont en lien avec l'émotivité de la personne (ex. neuroticisme), nous avons décidé d'implémenter dans notre algorithme prédictif une mesure de l'information émotive contenue dans les commentaires de l'utilisateur. Par exemple, Jean et al. [1] postulait l'existence d'une corrélation négative entre l'expression du sentiment d'angoisse dans les commentaires et le score à l'échelle Extraversion. Notre analyse des sentiments a été réalisée à l'aide de l'algorithme Roberta Base Go Emotions (disponible publiquement sur la plateforme Hugging Face), un modèle dérivé de l'architecture BERT (*Bidirectional Encoder Representations from Transformers*) développée par Google [4]. BERT s'appuie sur l'implémentation d'un *transformer* pour encoder l'information contenue dans le matériel textuel en préservant son contexte d'origine. Il est donc plus puissant que de simplement attribuer une valeur numérique à chaque mot. Après passage dans le *transformer*, le même mot peut se voir assigner une valeur complètement différente par BERT en fonction des mots qui l'entourent. C'est cette manière de préserver l'information qui

rend BERT aussi puissant. À la différence d'un encodage séquentiel, BERT est également bidirectionnel, ce qui signifie que peu importe si la séquence est lue de gauche à droite ou de droite à gauche. BERT lit la séquence des deux côtés à la fois. Bien que notre corpus soit principalement anglophone, cette fonctionnalité aurait pu être intéressante si nous avions eu à traiter des langages qui encodent le sens de droite à gauche comme l'hébreu ou l'arabe. À l'origine, BERT a été entraîné pour deux tâches (prédiction de mots masqués et prédiction de phrases adjacentes), cependant des versions comme RoBERTa ont simplifié cette approche. Il est donc facile de changer les étiquettes et la tâche et de lui faire évaluer, par exemple, un sentiment. Plus spécifiquement, ce modèle en particulier encode à l'aide du *transformer* un message pour ensuite inférer un score représentant la proportion du message (sur un intervalle $[0, 1]$) correspondant à chacune des 28 dimensions émotives; la somme des dimensions totalisant toujours 1. Les dimensions émotives choisies sont celles décrites dans le corpus de données GoEmotions sur lequel ce modèle en particulier a été entraîné. Ce corpus, également issu de Reddit, a été annoté par des humains, ce qui nous donne confiance en la qualité de son évaluation de l'émotion écrite. Nous pensons que la source similaire de données nous permettra d'avoir une évaluation très précise des émotions exprimées. Il est à noter que plusieurs de ces dimensions sont très proches et présentent un certain niveau de superposition. Par exemple, on peut s'attendre à ce qu'un message présentant un score élevé à l'attribut $grief$ présente également un score élevé à l'attribut *sadness*, de par la proximité conceptuelle entre ces étiquettes.

À la suite de la production de ces données émotionnelles de commentaires, les réseaux de neurones convolutionnel sont employés pour le traitement de ce signal. Les couches séquentielles de convolution suivie de la couche AdaptiveAvgPool1d permet de compacter le signal en un vecteur de taille fixe, qui est ensuite passé au travers de couches pleinement connectées pour produire une sortie de taille 5 (représentant les scores OCEAN).

D'autres réseaux de neurones basés sur les couches de convolutions ont été bâtis d'une façon à ce qu'ils soient capables d'accepter en entrée les données de signal émotionnelles et les données tabulaires. On pourrait ainsi s'attendre à obtenir de meilleures performances avec toutes les données utilisées en même temps, puisque différents liens pourraient être établis entre les données tabulaires et les données de signal.

Algorithme de prédiction par profil

S'est ensuite posé la question de quel algorithme choisir pour prédire les scores OCEAN à partir du profil. Au fil de nos essais (RandomForestRegressor, Ridge, KNeighborsRegressor, MLPRegressor, GradientBoostingRegressor), il s'agit des algorithmes *Random Forest*, ou forêt aléatoire, qui se sont révélés être les plus efficaces. Random Forest est un algorithme d'apprentissage machine supervisé du package SciKit-Learn. Son principe repose sur la création de plusieurs arbres de décision pour former un seul modèle (comme une forêt). Cette structure est plus robuste et moins sensible au surentraînement qu'un arbre seul. Le principal problème est qu'un arbre de décision unique peut apprendre les données d'entraînement sans être en mesure de généraliser ses apprentissages. *Random Forest* tire sa force de la diversité des arbres qui viennent, par accumulation, couvrir une plus grande partie de l'espace de possibilités. Il s'agit donc d'un compromis entre exploitation et exploration.[5]

Random Forest a été choisi pour plusieurs raisons. Tout d'abord, empiriquement, il est l'algorithme qui donnait les meilleurs résultats. Ensuite, au niveau conceptuel, *Random*

Forest était particulièrement adapté à la tâche. *Random Forest* est capable de traiter efficacement une large gamme de types de données sans nécessiter de normalisation poussée, contrairement à d'autres algorithmes comme les réseaux de neurones. Ça en faisait un candidat idéal pour les profils. De plus, il est robuste à la présence de NaN, et les profils en contenaient une grande proportion. Ensuite, les données PANDORA contenaient du bruit, comme constaté précédemment [1]. *Random Forest*, grâce à son approche par ensemble d'attributs, est robuste face à ce type de problèmes [5]. De plus, il permet d'évaluer l'importance des différentes features, ce qui est utile pour comprendre quels aspects du profil influencent le plus la prédiction des traits de personnalité.

Plusieurs ajustements ont été cependant été nécessaires pour optimiser notre implémentation de *Random Forest*. Notamment, il a fallut réaliser une optimisation des hyperparamètres, effectuée à l'aide d'un *grid search*. Les paramètres ajustés incluaient le nombre d'arbres ($n\_estimators$), la profondeur maximale des arbres (max_depth), ainsi que des critères comme $min\_samples\_split$ et $min\_samples\_leaf$ pour éviter les splits non informatifs et limiter le surajustement.

Algorithme de prédiction par commentaires

Enfin, nous avons dû travailler à choisir notre manière d'approcher le traitement des données textuelles des commentaires. Vu la quantité de calcul requise pour traiter 17 millions de commentaires avec une approche sensible au contexte comme TF-IDF, nous nous sommes plutôt tournés vers une méthode plus simple utilisant l'approche du *Bag of Words*, ou sac de mots. Il s'agit d'une manière de traiter les mots indépendamment de leur contexte dans la phrase et dans leur message pour simplement s'intéresser à leur présence, et si oui, à leur fréquence. Au cours des procédures de tests, différents sous-ensemble de taille variable ont été utilisés. Le sous-ensemble des N mots ($10 \leq N \leq 200$) dont le taux d'utilisation était le plus fortement corrélé avec les scores OCEAN était utilisé. Notre première implémentation d'algorithme en lien avec BOW était dans la démarche de génération des MBTI manquants. Pour prédire les scores MBTI, nous utilisons une approche fondée sur l'analyse textuelle, où chaque dimension (par exemple, Introversion/Extraversion) est traitée indépendamment par un modèle dédié. Ces modèles fonctionnent de la même façon, utilisant sur une représentation vectorielle des mots les plus discriminants entre les deux pôles d'une dimension donnée. Dans un premier temps, un échantillon des messages est partitionné en deux groupes en fonction de la dimension MBTI considérée. Par exemple, pour la dimension Introversion (I) vs Extraversion (E), nous séparons les messages des utilisateurs classés comme Introvertis de ceux classés comme Extravertis. Ensuite, un décompte de chaque mot permet d'identifier les mots les plus caractéristiques de chaque groupe. Ces mots, qui reflètent les caractéristiques lexicales importantes du profil, sont ensuite fusionnés en un ensemble unique, formant ainsi le vocabulaire de référence pour cette dimension.

Pour chaque utilisateur dont on souhaite prédire le score MBTI, nous construisons un vecteur de caractéristiques (*feature vector*) basé sur ce vocabulaire. Le vecteur encode le nombre d'occurrences de chaque mot significatif dans les messages de l'utilisateur. Par exemple, si le mot "*party*" est un terme discriminant pour la dimension I/E, sa fréquence dans les textes de l'utilisateur contribuera directement à la prédiction de son score sur cette dimension. Chaque modèle est ensuite entraîné à partir de ces vecteurs, associés aux étiquettes MBTI connues (issues des données d'entraînement). Une fois entraîné, le modèle peut prédire, pour un nouvel utilisateur, sa tendance sur la dimension cible en analysant la distribution des mots discriminants dans ses messages. Le score S de l'utilisateur, pour

un vecteur V de mots où chaque mot $V[i]$ possède un poids w_i est donc déterminé par l'équation suivante:

$$S = \frac{\sum_{i=0}^{|V|-1} w_i \times V[i]}{|V|}$$

Cette approche permet une prédiction fine et interprétable, puisqu'il nous est ensuite possible de consulter la liste et voir exactement quels mots ont été déterminants dans la classification, que ce soit car le poids de cet indice est important, ou car l'élément du vecteur avait une taille importante.

Lorsqu'est venu le temps de réaliser la prédiction des dimensions OCEAN à l'aide des messages, nous avons utilisé une version modifiée de cet algorithme pour prendre en compte le fait que les scores OCEAN sont donnés en percentiles alors que l'algorithme MBTI utilisait un classificateur binaire. Nous avons également constaté que la précision des prédictions différait beaucoup entre les différentes dimensions. Pour aider l'algorithme, nous avons donc réalisé un pipeline utilisant les scores inférés aux dimensions précédentes pour augmenter la précision des dimensions avec la plus grande erreur. Par exemple, l'échelle *Openness* ayant un R^2 de 0.68, nous commençons par cet attribut. Le suivant recevra comme entrée le vecteur de mots, mais également la prédiction *Openness*. Ces ajouts sont faits itérativement jusqu'à ce que la dernière échelle utilise 4 scores prédits et le vecteur de mots.

Procédure itérative de tests

Notre approche de conception du *pipeline* s'inscrivant dans un cadre de développement itératif, nous avons eu besoin d'une méthode standardisée d'évaluation de nos tests pendant le développement des modèles prédictifs. L'objectif était double: d'abord, éviter de devoir déterminer au cas par cas les métriques d'évaluation de chaque test, et ensuite, encapsuler la procédure afin de la rendre accessible aux collaborateurs moins familiers avec l'entraînement de modèles d'apprentissage automatique. Pour ce faire, nous avons réalisé deux fonctions différentes ($fit()$ et $evaluate()$) capables de prendre en entrée un modèle (pré-entraîné ou non) et de gérer automatiquement les spécificités liées à son architecture, telles que la gestion des valeurs manquantes (NaN) ou la nature des prédictions (multiples ou uniques). La fonction $evaluate()$ possède également un argument optionnel, permettant de lui passer un second modèle avec lequel comparer le premier en termes de performances prédictives. Cet outil a été crucial dans la sélection des modèles à intégrer au pipeline final. Ces méthodes permettent également d'utiliser différents échantillonnages d'auteurs pour assurer la validité de notre processus de validation. Ces échantillons, entraînement (n=1248), test (n=160) et remise (n=160) avaient pour but de s'assurer que les algorithmes étaient réellement évalués sur des données qu'ils n'avaient jamais vu.

L'utilisation systématique d'une même méthode d'évaluation a toutefois soulevé la question du choix des métriques à calculer. Nous avons retenu l'erreur moyenne (RMSE), celle-ci correspondant au métrique d'évaluation final de nos prédictions, ainsi que le coefficient de détermination (R^2) afin de quantifier la proportion de variances expliquées par le modèle. Plus précisément, en lien avec le RMSE, nous avons analysé séparément l'erreur de chaque score OCEAN, sous la forme d'une liste de *float* (par exemple, [30.24511261 32.12280764 31.23772002 31.89197057 31.35113002] (respectivement [O, C, E, A, N])). Cependant, nous rapporterons ici une mesure singulière du RMSE à des fins de lisibité, calculée en faisant la moyenne des 5 RMSE. Pour les classificateurs binaires, nous avons également inclus une

matrice de confusion, permettant de calculer le rappel et la précision.

Notre première implémentation de la procédure de test a été de réaliser un modèle prédictif uniquement basé sur la médiane. Ce test avait pour but d'établir une référence contre laquelle comparer nos modèles suivants pour la suite de nos tests. Pour chacune des dimensions, indépendamment des caractéristiques de l'utilisateur, le score OCEAN retourné correspondait au score médian dans l'échantillon. Sans grande surprise, les statistiques descriptives de ce test se sont révélées très faibles. L'erreur moyenne était de 34.1 et le coefficient $R^2 \approx 0.06$. Autrement dit, la médiane ne prédit qu'une infime partie du score général de l'utilisateur, et l'erreur moyenne tend vers l'erreur d'une prédiction aléatoire sur l'intervalle [0,100]. Nous n'utiliserons donc pas ce modèle dans le *pipeline* de prédiction final, mais il permet d'établir un seuil de comparaison général (Table 1).

Notre seconde implémentation de la procédure de test a été de réaliser un modèle prédictif des scores MBTI des utilisateurs. Comme expliqué dans notre article précédent, bien que les données du jeu de données PushShift révélaient une corrélation intéressante entre les scores MBTI et OCEAN, une proportion significative d'utilisateurs n'avaient pas déclaré de score MBTI. Dans l'impossibilité d'inférer les scores manquants à partir des scores OCEAN, nous avons dû nous appuyer sur le contenu textuel des commentaires des utilisateurs. En utilisant l'approche BOW, nous avons entraîné quatre modèles différents (un par échelle du MBTI). Le fonctionnement précis de la transformation des messages est détaillé dans la section dédiée aux algorithmes utilisés. Différentes architectures ont été tentées soit, la régression linéaire, la forêt aléatoire (*Random Forest*) et le *boosting* par gradient. Dès les premiers tests, la méthode de boosting par gradient a démontré des performances supérieures. Après confirmation de sa viabilité sur les différentes dimensions MBTI (avec des mesures de précision comprises entre 0,7 et 0,8 sur un échantillon initial restreint (N=400)) nous avons procédé à l'entraînement final sur un échantillon élargi (N=2500). Ce modèle optimisé a ensuite été utilisé pour inférer les scores MBTI manquants dans notre jeu de données complet.

Notre troisième implémentation de la procédure de test a été de réaliser un modèle prédictif uniquement basé sur les informations contenues dans les profils d'utilisateur. Pour ce faire, nous avons évalué plusieurs architectures d'apprentissage automatique, incluant la régression linéaire, la forêt aléatoire (*Random Forest*) et les réseaux de neurones profonds. Toutefois, les données initiales présentaient un niveau de bruit trop élevé pour permettre une prédiction efficace, comme en témoignait la valeur négative du coefficient de détermination ($R \approx -0,05$) obtenue lors des premiers entraînements. Pour cette raison, nous avons réalisé un second prétraitement des données (voir article précédent), et avons bonifié les données à l'aide de nouveaux attributs, comme le nombre de jours d'utilisation de Reddit, ainsi que la quantité moyenne de messages par jour. Les entraînements réalisés sur les données nettoyées se sont révélés beaucoup plus encourageants, avec une erreur moyenne en dessous de notre *baseline*. Les modèles de type *Random Forest* se sont particulièrement distingués, affichant une précision supérieure aux autres approches. Une optimisation fine de leurs hyperparamètres (voir Table 1) nous a permis d'identifier une configuration optimale minimisant l'erreur de prédiction.

Lors d'analyse subséquente, nous avons déterminé que la performance globale de ce modèle était beaucoup impactée par l'hétérogénéité des performances sur les sous-échelles de OCEAN. Par exemple, alors que l'échelle d'extraversion présentait un $\mathbf{R^2}$ élevé (0,68), celle d'ouverture affichait une valeur plutôt faible (0,25), suggérant une capacité prédictive inégale. D'autres méthodes ont donc été ensuite développées pour prédire avec plus de précision

les échelles dont les métriques *Random Forest* laissaient à désirer.

Notre quatrième implémentation de la procédure de test a été de réaliser un modèle prédictif uniquement basé sur les informations contenues dans les commentaires de l'utilisateur. Pour ce faire, nous avons évalué plusieurs architectures nommées précédemment, mais c'est l'architecture GradientBoostingRegressor qui a été en mesure de mieux gérer le vecteur BOW décrit dans la section **Sélection des Algorithmes - Algorithme de prédiction par commentaires**. Les résultats initiaux, principalement au niveau de la prédiction des scores MBTI se montraient suffisamment convaincants, avec un coefficient de détermination correct ($R \approx 0.2$), mais un excellent rappel (0,99) et une bonne précision (0,8). Ces résultats initiaux sur le MBTI nous ont donc encouragés à étendre l'implémentation à la prédiction des scores OCEAN. Lorsque nous avons réalisé son implémentation, cependant, nous avons rapidement constaté que les différentes dimensions OCEAN souffraient de grandes disparités dans les performances de leurs modèles respectifs. L'écart maximal entre le RMSE de deux échelles était de 8,67. Nous avons donc modifié notre manière de prédire les scores OCEAN en gradant la précision de chaque modèle individuellement, et en itérant du plus fiable vers le moins fiable en lui fournissant également, non seulement le vecteur de mots, mais également les valeurs OCEAN prédites pour les dimensions précédentes. En utilisant cette technique, l'écart maximal entre le RMSE de deux échelles a diminué pour se retrouver à 4,96. Nous pensons que cette démarche est un pas dans la bonne direction, mais soutenons l'idée que certaines dimensions OCEAN semblent plus difficiles à inférer que d'autres en se basant uniquement sur les commentaires de l'utilisateur.

Notre implémentation finale de la procédure de test à été de réaliser un modèle prédictif basé sur l'analyse des sentiments, prenant en entrée à les données issues de l'algorithme **Roberta Base Go Emotions**. Dès le départ, son implémentation a été difficile en raison de la quantité limitée d'algorithmes en mesure de prendre en compte adéquatement ce type de données (voir section **Architecture finale du système** - Figure 3). Nous avons tenté d'entraîner un réseau de neurones profonds (RNN) sur ceux-ci, car cette architecture nous semblait la plus à même de capturer la complexité des données, de même que leur nature évolutive dans le temps. Malgré les efforts de l'équipe, l'entraînement tendait à empirer l'erreur de mesure pour la majorité des échelles OCEAN. Par exemple, après 100 époques d'entraînement, le RMSE moyen des échantillons de validation était de 29.9862, alors qu'après 1000 époques, le RMSE moyen a augmenté jusqu'à 36.281 (Fig. 1). N'étant pas en mesure de réaliser un algorithme prédictif utilisant ce type de données qui ai un RMSE inférieur à notre mesure baromètre, nous avons donc décidé de ne pas inclure cet algorithme dans notre *pipeline* final. Il est à noter que cette technique a cependant donné des résultats prometteurs sur certaines échelles, notamment *Neuroticism et Agreeableness*. Notre hypothèse est qu'il s'agit des dimensions OCEAN les plus explicitement corrélées aux émotions des utilisateurs. Des études subséquentes sur la viabilité de cette technique seront nécessaires.

Enfin, vous trouverez ci-après un tableau comparatif des erreurs de mesure (RMSE) pour chacun des tests, en fonction des modèles utilisés. Pour faciliter l'interprétation, chacun est mesuré en termes de pourcentage d'amélioration par rapport au modèle de base (prédiction par médiane).

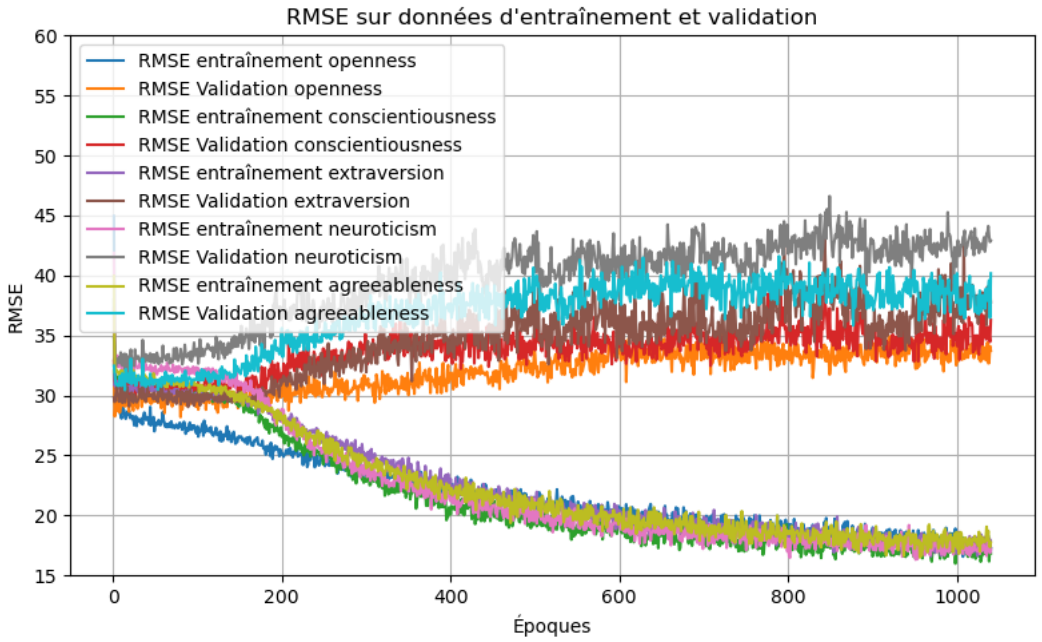

Figure 1. Exemple d'évolution de l'erreur de mesure pendant l'entraînement d'un réseau de neurones profonds sur les données d'analyse de sentiment.

Type de test	Modèle	Erreur moyenne	Amélioration (%)
Test MBTI	XGBRegressor	0.1	N.D.
Test Médiane	Custom	34.1	Baseline
Test de prédiction par profil	RandomForestRegressor	28.61	16.09%
	Ridge	28.83	15.45%
	KNN	30.28	11.2%
	KNN (custom weights*)	31.35	8.06%
	FCNN	30.54	10.43%
	CNN	31.31	8.05%
Test de prédiction par commentaire	RandomForestRegressor	30.27	11.23%
	Ridge	30.20	11.43%
	KNN	31.35	8.06%
	KNN (custom weights*)	30.4	10.88%
	FCNN	31.36	8.06%
	CNN	31.23	8.12%

Table 1. Métriques d'évaluation de différent modèles testés

Études de cas

Nous analyserons ici de manière détaillée certains cas concrets d'utilisateurs ayant contribué à informer notre prise de décision dans le développement de notre algorithme. Les utilisateurs ont été sélectionnés pour couvrir une diversité de profils et de comportements, permettant ainsi de donner un aperçu des forces et des difficultés de notre modèle.

L'utilisateur AbstractStateMachine est un des utilisateurs pour lequel nos prédictions ont été, en général, dans la norme attendue par notre erreur moyenne ($\pm 4,505$ points autour

de la valeur attendue). Cela peut s'expliquer par le fait que cet utilisateur avait un nombre relativement restreint de messages à partir desquels inférer son score OCEAN (2045 au total), que ceux-ci étaient relativement courts, sans prise de position assumée, et aussi car son profil PANDORA ne comportait pas de scores MBTI. Nous avons donc dû inférer ce dernier. Il est possible qu'une erreur dans l'inférence d'une ou deux échelles du MBTI ($p \geq 0,59$ considérant la précision de notre algorithme) soit venue interférer par la suite avec nos différents algorithmes et a contribué à introduire du bruit dans le système. Nous maintenons cependant que cette étape était nécessaire. L'erreur moyenne de mesure pour AbstractStateMachine était en effet 6 points plus élevés lorsque nous n'utilisions pas de scores MBTI, plutôt que les scores prédits. Pour cette raison, nous pensons qu'il reste pertinent de continuer à utiliser l'algorithme de prédiction des scores MBTI, et ce même s'il a de grandes chances de se tromper sur un ou plusieurs aspects. Nous estimons que le gain en information justifie son inclusion dans le *pipeline*.

L'utilisatrice Frenchitwist a été l'utilisatrice pour laquelle nos prédictions ont été le plus précises et ce à la fois pour la prédiction par profil, par commentaires, et par comité. Cela peut s'expliquer par plusieurs facteurs. Premièrement, il s'agit d'un des utilisateurs dont le profil PANDORA était le plus complet (nombre minimal de NaN). Nous avions donc une bonne quantité d'information sur laquelle nous appuyer pour inférer ses scores OCEAN à partir de son profil. Il est également possible que, contrairement à d'autres utilisateurs, ces informations souffraient également d'un bruit minimal. Ensuite, cet utilisateur possédait une quantité adéquate de messages, dans la moyenne supérieure, mais sans bruiter l'entrée en produisant un volume trop conséquent de messages de faible qualité. On voit, en effet, que la longueur moyenne des messages de Frenchitwist est de 158.78 caractères. Finalement, ce qui a probablement beaucoup aidé est que les scores aux échelles de personnalité de cet utilisateur sont relativement proches de la médiane, et ce pour la plupart des dimensions. Celà fait en sorte de minimiser l'erreur de mesure puisque même si l'algorithme sur ou sous-évalue le score, l'erreur moyenne sera minimale. Voyant ces caractéristiques, Frenchitwist a également servi d'individu-test lors de certaines explorations pour valider la précision d'un algorithme lorsqu'il possède la quantité maximale d'information de bonne qualité sur un utilisateur.

L'utilisateur Strongbhoy a été l'utilisateur pour lequel nos prédictions ont été les plus disparates, mais également les plus fautives. La grande variabilité des scores qui lui ont été assignés peut être expliquée par plusieurs facteurs, mais nous ne pouvons pas conclure avec certitude que l'un d'entre eux est réellement la cause. Il se pourrait par exemple que les informations fournies par l'utilisateur sur son profil soient erronées, que ce soit intentionnel ou pas. Il est également possible que la personne opérant ce compte, si on en croit la faible quantité de commentaires rédigés, mais sur une longue période, l'utilise comme *throwaway account* et donc présente une façade au reste de la communauté qui n'est pas exactement en accord avec sa personnalité authentique. Enfin, il est aussi possible que cet utilisateur éprouve des problèmes de santé mentale tels qu'un trouble de l'humeur ou de la personnalité et que la grande variabilité que l'on observe en ligne soit en fait le signe d'une personnalité profondément instable dans la réalité [6]. Le fait est que cet utilisateur, en plus d'avoir des scores relativement extrêmes aux différentes échelles OCEAN, ne semble pas être dans les extrêmes pour les dimensions censées être corrélées. Étant incapables de déterminer avec certitude la cause de l'erreur et son extrême variabilité, nous avons choisi de ne pas implémenter de solution, de manière à ne pas introduire de biais contre cet utilisateur ou d'autres utilisateurs présentant un profil désorganisé.

L'utilisateur -Fwer a été identifié comme étant l'utilisateur le plus objectivement difficile à prédire dans notre échantillon, en raison de ses données de profil éparses et bruitées, ainsi que son activité limitée. Il n'a publié qu'un seul message, composé d'un copier-collé de son résultat au test "Understand Myself" suivi d'un commentaire personnel sur celui-ci.

> October 6, 2017 Results via understandmyself.com Agreeableness 46 Compassion 16 [...]. 18 yo male from Australia.Very confident that these results reflect my personality. [...] Being extremely high in orderliness, I'm the type who can't function without a proper routine, but, if I fail to implement it perfectly, it all goes to shit. I SPENT MORE THAN 6 HOURS PLAYING MADDEN YESTERDAY AND I DON'T EVEN ENJOY PLAYING IT. Would appreciate any other input :)

De par les contraintes reliées au projet, nous n'étions pas autorisés à extraire directement les scores rapportés dans de ce type de commentaire, le rendant fonctionnellement inutilisable, sinon pour la seconde partie. Nous avons donc dû nous appuyer presque entièrement sur les méthodes de prédiction basées sur le profil pour cet utilisateur. Nous sommes conscients que même les dimensions MBTI qui lui ont été assignées risquent d'être inexactes et d'avoir entraîné une mauvaise prédiction quant aux scores OCEAN.

L'utilisateur neketa1 a également soulevé un défi important dans l'analyse de ses messages, rendant la prédiction de ses scores plus ardue. Dans la rédaction de ses messages, celui-ci utilisait une forme d'enluminure numérique communément appelée Zalgo (Fig. 2). Cette technique utilise la capacité d'une boîte de texte à superposer différents niveaux de textes pour créer un effet visuel utilisant différents caractères générés aléatoirement pour donner l'impression que le message initial souffre d'une corruption des données. Le résultat encode bel et bien le message de base, mais ajoute énormément de bruit autour de celui-ci lorsque ce message est stocké sous forme de String comme c'était le cas dans le corpus de données PushShift.

Figure 2. Exemple d'utilisation de Zalgo

Encodage du même message en String: Armageddon¡@?&&?‘‘(!"; [...]
La solution que nous avons trouvée pour nous permettre de traiter ce type de messages est de procéder à la séparation des caractères alphanumériques d'avec les caractères reliés à la ponctuation. De cette manière, Armageddon est traité indépendamment, et comme la suite de caractères ajoutés ne contient pas de lettres, celle-ci est éliminée lors du processus de lemmatisation. Évidemment, il nous aura fallu revisiter notre manière de réaliser notre prétraitement des données pour implémenter cette solution.

Architecture finale du système

L'architecture finale retenue est un *pipeline* en trois étapes divisé de la façon suivante:

(1) Prétraitement des données du profil et des commentaires
(2) Analyse des données et production des prédictions intermédiaires
(3) Mise en commun et réalisation de la prédiction finale

1) Le pré-traitement des données du profil est réalisé en isolant les attributs sélectionnés (author, MBTI, OCEAN, age, gender, is_female, num_comments, active_days, comments_per_day), normalisant les attributs numériques, et transformant les attributs nominaux en attributs numériques.Le pré-traitement des données des commentaires est réalisé en isolant les attributs sélectionnés (author, body et timestamp), normalisant les attributs numériques, et réalisant une série de transformations sur le corps du texte pour en faciliter l'analyse. Ces transformations incluent la séparation de la ponctuation, l'uniformisation de la casse, le retrait des mots vides et la lemmatisation du texte.

2) L'analyse des données est réalisée à l'aide d'une succession d'étapes destinées à extraire l'information contenue dans le corpus et suppléer l'information manquante.

- Détermination des mots significatifs pour chaque dimension
- Calcul des vecteurs reliés à chaque utilisateur
- Inférence du MBTI (si manquant)
- Prédiction intermédiaire basée sur le profil
- Extraction des scores de sentiment de chaque message
- Classement des émotions pour chaque utilisateur
- Prédiction intermédiaire basée sur les messages

Les utilisateurs sont divisés en différents groupes en fonction de la quantité de données disponibles sur eux. Cette division est réalisée en ajoutant un attribut pour les utilisateurs dont le profil est épars, et une dimension pour les utilisateurs dont la quantité de messages est en dessous d'un seuil prédéterminé.

3) Un modèle de prédiction par comité utilise les différents scores prédits pour réaliser une prédiction finale sur le score de l'utilisateur à cette dimension.

L'enchaînement exact des différentes étapes et les flux d'information sont représentés graphiquement dans la figure 3 ci-dessous.

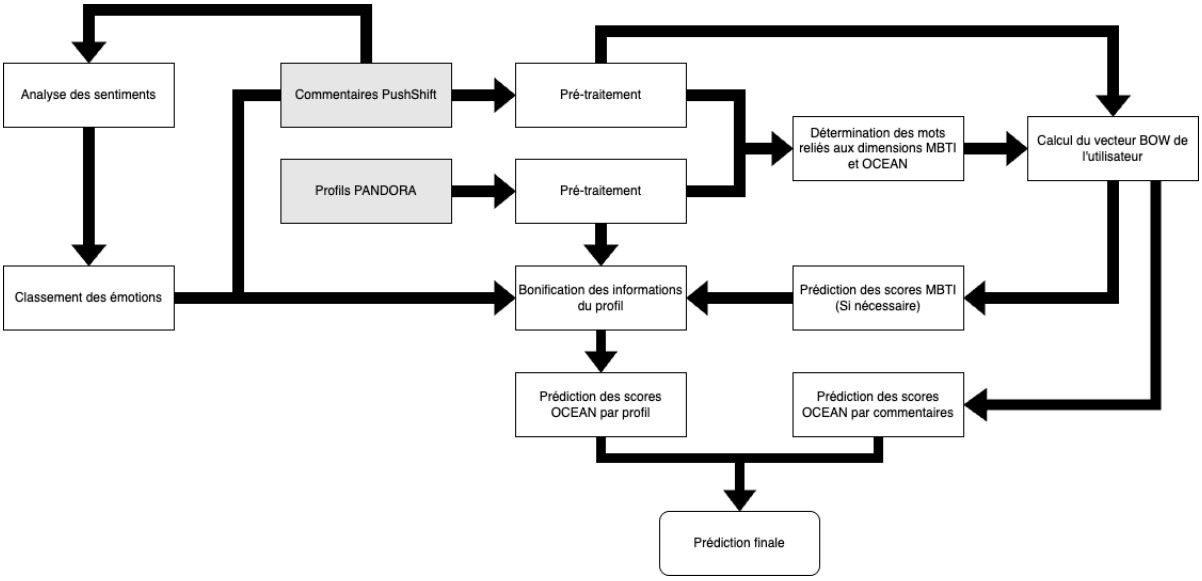

Figure 3. Architecture finale l'algorithme prédictif

Conclusion

Dans le cadre de ce projet, nous avons développé un algorithme prédictif capable d'estimer les scores aux dimensions de personnalité OCEAN d'un utilisateur à partir de son profil et de ses commentaires Reddit. Cette avancée représente à elle seule une réussite. En nous appuyant sur des techniques de NLP, nous avons amélioré sa précision par rapport à notre mesure baromètre. Malgré tout, nous pensons qu'il laisse place à beaucoup d'amélioration et sa performance n'égale pas encore celle des algorithmes actuels. Notre approche s'inspire des dernières publications scientifiques en matière de psychométrie et s'appuie sur des processus rigoureux de validations par test, garantissant une certaine robustesse et la pertinence des résultats.

Conscients des enjeux éthiques liés à ce type d'analyse, nous avons été en mesure de tenir compte des biais potentiels d'un tel système pendant tout le cycle de développement. Une attention particulière a été portée aux risques de discrimination, notamment envers les communautés vulnérables (comme les personnes trans ou celles souffrant de troubles de santé mentale). Bien que les données initiales présentent des limites et biais issus du monde réel, nous avons mis en place des garde-fous en équipe pour minimiser l'impact négatif de notre système.

Une des leçons les plus importantes que nous avons apprise au cours de ce projet est l'importance de la classe de complexité des algorithmes choisis pour réaliser le traitement des données massives. Lorsque l'algorithme doit être appliqué sur 17 millions de messages, la moindre inefficacité dans le traitement se transforme en dizaines d'heures de calcul en plus. Lors de notre choix de l'algorithme d'analyse de sentiment, nous n'avons pas été suffisamment conscients de ce point, ce qui a entraîné un temps approximatif de traitement de 100 heures à se répartir entre les différents membres de l'équipe pour traiter l'ensemble des commentaires. Pour pallier à ce problème, nous avons dû avoir recours au service de Cloud-Computing offert par HuggingFace pour réaliser les calculs sur leurs serveurs distribués, ce qui a réduit l'impact de notre côté. Il reste qu'un algorithme plus performant aurait été préférable.

L'analyse de sentiment a également été une grande source d'apprentissage au cours de ce projet. Nous étions convaincus de sa pertinence en lien avec la question de recherche, mais les résultats empiriques ne concordaient pas avec notre intuition. Nous avons tenté de l'implémenter de différentes manières, que ce soit à l'aide de *TimeSeries* ou d'analyse de moyennes. Malheureusement, aucune de ces méthodes n'a réussi à montrer une capacité de distinction intéressante, ne serait-ce que sur une seule dimension de OCEAN. L'entraînement de modèles sur ces données, non seulement échouait à trouver un optimum local minimisant l'erreur moyenne, mais empirait l'erreur de prédiction dans l'échantillon de validation (Fig. 1).

Nous avons également été confronté à de grandes difficultés à extraire des informations des commentaires des utilisateurs. Nous étions conscient des défis que ce type de données représentaient, mais nous avions sous-estimé sa complexité. Il est possible qu'un étiquettage d'uniquement 1500 utilisateurs ai nuit à notre capacité à entraîner adéquatement un modèle capable d'extraire les features les plus importantes des commentaires. Le langage naturel étant naturellement épars et très bruité, il n'est pas surprenant que très peu de lien significatif entre celui-ci et la personnalité des utilisateurs ai été découvert.

Ces déboires d'implémentation nous ont appris à rester fluides et créatifs. Éventuellement, nous avons opté pour une implémentation simple, par classement, et cette approche a éventuellement présenté des métriques justifiant de l'utiliser dans notre architecture finale. Nous avons donc constaté que parfois les meilleures solutions ne sont pas nécessairement les plus mathématiquement complexes.

Mentions

L'équipe de travail déclare ne pas être en situation de conflit d'intérêts. Cette étude est le travail original de ses auteurs et n'a pas été produite en collaboration avec quelque organisation tierce pouvant bénéficier de ses résultats. Les auteurs n'ont été dédommagés d'aucune façon pour leur travail. L'étude respecte également les lignes directrices en matière d'utilisation de l'IA générative définies pour le cours GLO-4027.

References

[1] A. Jean, J. Carré, B. Diallo, and O. Bocoum. "Analyse préparatoire à la prédiction des profils OCEAN d'utilisateurs de Reddit". In: *Proceedings of the Canadian Conference on Artificial Intelligence* (2025).

[2] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. "The Pushshift Reddit Dataset". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (2020), pp. 830–839. DOI: 10.1609/icwsm.v14i1.7347. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7347.

[3] M. Gjurković, V. M. Karan, I. Vukojević, M. Bošnjak, and J. Snajder. "PANDORA Talks: Personality and Demographics on Reddit". In: *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Ed. by L.-W. Ku and C.-T. Li. Online: Association for Computational Linguistics, June 2021, pp. 138–152. DOI: 10.18653/v1/2021.socialnlp-1.12. URL: https://aclanthology.org/2021.socialnlp-1.12/.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[5] A. Parmar, R. Katariya, and V. Patel. "A review on random forest: An ensemble classifier". In: *International conference on intelligent data communication technologies and internet of things (ICICI) 2018*. Springer. 2019, pp. 758–763.

[6] O. Friborg, E. W. Martinsen, M. Martinussen, S. Kaiser, K. T. Øvergård, and J. H. Rosenvinge. "Comorbidity of personality disorders in mood disorders: a meta-analytic review of 122 studies from 1988 to 2010". In: *Journal of affective disorders* 152 (2014), pp. 1–11.