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

Analyse préparatoire à la prédiction 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 analyse les activités en ligne et les profils d'utilisateurs de Reddit pour concevoir une méthodologie permettant de prédire leurs scores au test de personnalité OCEAN (Big Five). En utilisant 17 millions de commentaires collectés via l'API Pushshift et le profil des utilisateurs associés, nous visons à extraire le contenu sémantique (selon une approche *Bag of Word* (BoW)) et le sens des métadonnées de l'utilisateur pour entraîner un algorithme prédictif. Nous proposons un *pipeline* de traitement des données qui comprend une réduction de la dimensionnalité, le prétraitement du texte, l'inférence des scores MBTI et l'analyse des sentiments afin de prédire le score avec exactitude. En fonction de la quantité de données disponible sur l'utilisateur, différents traitements sont réalisés dans le *pipeline*. Notre méthodologie inclut une validation par retenue, justifiée par le nombre limité de profils OCEAN complets disponibles.

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.[1] 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.[2] sur ce corpus, nous cherchons à estimer les scores de personnalité OCEAN des utilisateurs pour qui ceux-ci sont omis. Nous élaborerons ici une méthodologie visant à prédire ces scores en se basant sur les données contenues dans les deux corpus.

Propriétés statistiques

Notre corpus de données est composé de 2 tables de données. Les tables feront l'objet d'analyses séparées.

Analyse des profils

Notre corpus de données est composé de 10,295 profils d'utilisateurs Reddit. Chaque profil peut posséder jusqu'à 38 attributs. Cependant, $62,215\%$ des données hypothétiques des profils sont manquantes (NaN). Selon les auteurs de l'article, ce serait car les profils ont été générés automatiquement à partir des données soumises par les utilisateurs dans leur commentaires[2]. Nous avons considéré l'option de garder les valeurs NaN pour éviter d'introduire de fausses valeurs dans nos données. Cependant, agir ainsi rendait plusieurs attributs inutilisables dans le processus prédictif. En conséquence, nous opterons plutôt pour une approche générative de certains NaN lors du traitement des données (Section 4).

This article is © 2025 by author(s) as listed above. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (https://creativecommons.org/licenses/by/4.0/legalcode), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the author(s) identified above.

Le profil des utilisateurs contient à la fois leurs informations personnelles (âge, genre, sexe, nationalité) et psychométriques (MBTI, Ennéagramme, profil OCEAN). L'utilisateur moyen est un homme adulte (dans la mi-vingtaine) résidant aux États-Unis, parlant anglais, et ayant rédigé environs 606 commentaires (médiane) sur le site. Il est à noter que les attributs du profil ont été générés automatiquement à partir des commentaires des utilisateurs par les auteurs du corpus Pandora. Elles contiennent donc un bruit significatif, allant même jusqu'à des informations entièrement erronées, par exemple un utilisateur de décrivant comme «32M» dans un de ses commentaires est listé comme une femme de 18 ans dans les données du corpus.

En ce qui a trait aux données psychométriques, 9084 utilisateurs possèdent un profil MBTI, 1608 possèdent un profil OCEAN et 794 possèdent un ennéagramme.Vu la quantité de données disponible, nous estimons donc qu'il est plus probable que nous soyons en mesure de trouver des corrélations significatives entre le score OCEAN et les scores MBTI, du moins plus qu'avec l'ennéagramme. Il est à noter que dans seulement 1408 cas, le profil OCEAN rapporté par l'utilisateur est complet. Les autres possèdent une valeur manquante à au moins une sous-échelle. Dans notre jeu de données, 160 utilisateurs ont vu l'ensemble de leur score OCEAN caviardé (valeur de -1 à chaque attribut), nous donnant comme mission d'arriver à prédire les valeurs expurgées. Sans ces valeurs, la distribution échantillonnale des différentes échelles du profil OCEAN est la suivante (Fig. 1).

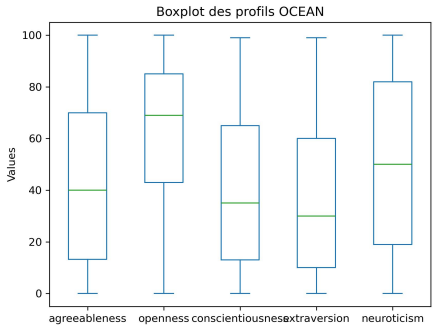

Figure 1. Distribution des scores OCEAN, séparés par échelle

Il est à noter que les scores OCEAN sont également soumis à un bruit important. Tout d'abord, car ils sont issus de scores rapportés volontairement par l'utilisateur, qui aura pu les falsifier en amont de la collecte de donnée. Les auteurs du corpus Pandora ont ensuite normalisé les scores au moyen d'un algorithme inconnu qui a pu introduire du bruit supplémentaire dans les valeurs rapportées. Cet aspect est important à garder en tête lors de l'analyse car ces dimensions serviront d'étiquette à notre algorithme de prédiction. Un mauvais étiquetage, s'il y a lieu, demanderait d'être compensé à l'intérieur du *pipeline*.

À des fins d'analyse exploratoire, nous avons réalisé une carte de chaleur présentant les corrélations des différents attributs du profil entre-eux (Fig. 2).La grande majorité des attributs présentent une corrélation presque nulle entre eux. Nous pourrons donc retirer certains attributs qui ne contribuent pas à la prédiction du score OCEAN. Le résultat semble également indiquer que certains groupes d'attributs sont corrélés très fortement entre eux, par exemple le genre et les attributs de géolocalisation. Ceux-ci pourront être unifiés sans perte d'information. Sans surprise, les tests sont fortement corrélés au type de résultats qu'ils donnent. Ce qui nous intéresse particulièrement dans ce graphique sont les attributs

les plus fortement corrélés avec les échelles du Big Five. On constate que les échelles du MBTI ont presque toutes une corrélation supérieure à 0.5 pour un attribut du Big Five. Nous pourrons donc les utiliser comme prédicteurs partiels du score OCEAN.

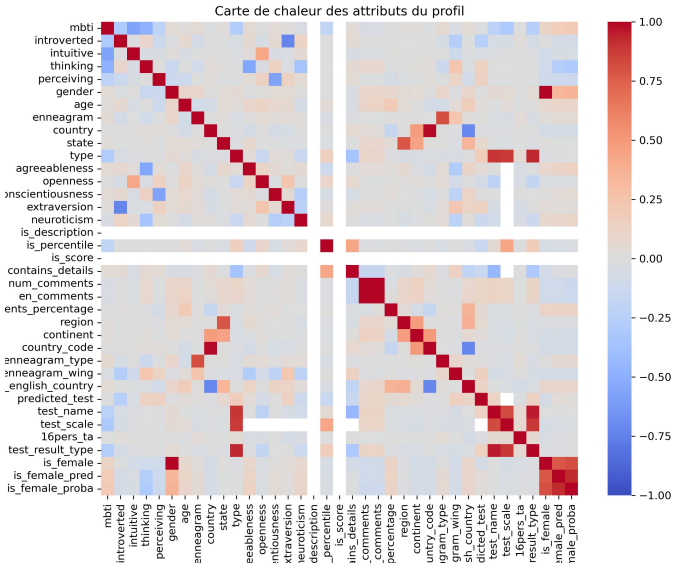

Figure 2. Carte de chaleur (*Heatmap*) des attributs du profil

Analyse des commentaires

En plus des profils, notre corpus de données rassemble 17,640,062 commentaires rédigés par ces mêmes utilisateurs. Chaque commentaire peut posséder jusqu'à 17 attributs. Les commentaires sont répartis dans 46214 subreddits différents, allant de la psychologie, aux jeux vidéos, en passant par l'actualité et la politique. En règle générale, la majorité des scores des commentaires tournent autour de 0, avec une petite quantité de commentaires possédant un score énorme (allant jusqu'à 56727). Cette répartition des scores représente bien la nature du site web, qui peut artificiellement mettre de l'avant certains commentaires si ceux-ci sont populaires au sein de leur subreddit original.

La quantité de messages par utilisateur suit une loi de puissance (Fig.3). Le premier quartile, en terme de nombre de messages envoyés, est d'environs 200 messages. Celà implique que pour une grande quantité d'utilisateurs, les scores OCEAN doivent être inféré à l'aide d'une quantité très restreinte de texte. Pour ce faire, nous avons la possibilité de plus appuyer notre prédiction sur les informations contenues dans le profil, ou bien de faire un traitement particulier des messages. Vu la corrélation très limitée entre les scores OCEAN et le reste des attributs du profil, nous opterons pour une approche spécialisée des commentaires (Section 4).

La quantité de mots dans le corps des commentaires varie énormément (selon une loi de puissances). Plusieurs commentaires ne possèdent qu'un seul mot, alors que le plus long commentaire contient une courte nouvelle avoisinant 10,000 mots. La fréquence d'utilisation

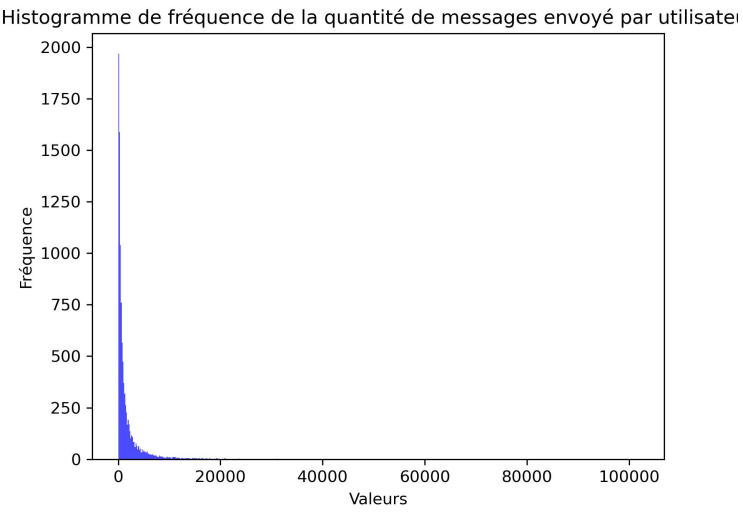

Figure 3. Distribution de fréquence du nombre de messages envoyés par utilisateur

des mots à l'intérieur des commentaires varie également selon une loi de puissance. Le mot le plus populaire, en accord avec les statistiques de la langue anglaise, est «*the*», suivi de différentes particules et déterminants. Par la suite, on retrouve les verbes usuels, puis enfin les différents noms propres et communs. Enfin, nous observons une longue queue de mots utilisés une ou deux fois. Certains sont des instances d'autres mots mal orthographiés, alors que d'autres sont des liens hypertextes, des omissions de tabulation, différentes graphies de smiley, etc. Nous pensons donc qu'il n'est pas pertinent de traiter une grande partie de la longue queue de cette distribution puisque l'information sémantique qu'elle porte est négligeable comparé au coût en ressource qu'elle requiert. Nous obterons pour une lemmatisation (Section 4) qui aidera à corriger certaines graphies fautives, mais nous pensons que la manière la plus simple de traiter la longue queue sera de l'uniformiser sous la forme d'un token unique facilement reconnaissable (ex.MOT_INCONNU).

Attributs sélectionnés

Les attributs sélectionnées à la suite de notre analyse statistique des données sont les suivants (Table 1). Il est à noter qu'il ne seront pas les seuls attributs utilisés comme prédicteurs de score OCEAN, mais ceux qui nous semblent présenter la plus forte probabilité empirique de contribuer à notre algorithme prédictif.

Profil	**Commentaire**
author	author
mbti	body
OCEAN	wordcount
gender	flair
age	subreddit
	score

Table 1. Attributs sélectionnés

L'attribut «author» a été choisi par défaut. Il s'agit de l'attribut-pivot entre les deux jeux de données, et il serait impossible de réaliser d'inférence psychométrique sans pouvoir relier les commentaires à un individu spécifique.

En ce qui a trait aux profils, nous avons choisi l'attribut «mbti» parce que les différentes échelles du MBTI présentent des corrélations intéressantes avec certaines échelles de OCEAN. Cependant, l'implémentation sera de retirer la colonne «mbti» des données et de ne conserver que les colonnes contenant les scores à ses différentes échelles. Le score final au MBTI peut facilement être reconstruit à partir des scores aux échelles. Nous avons choisi l'attribut «OCEAN», référant ici aux 5 échelles Big Five, parce qu'il s'agit de la cible de notre processus prédictif et les utilisateurs qui possèdent ce score nous permettront d'entrainer nos algorithmes prédictifs à l'aide de données réelles. Nous avons choisi l'attribut «gender» parce que les études en psychométrie tendent à constater qu'il existe des différences significatives de profil de personnalité entre les différents genres [3]. Malgré le manque de littérature scientifique à ce sujet, nous jugeons qu'il est également approprié d'inclure les personnes trans et non-binaires dans l'analyse puisqu'il nous apparait logique de penser que ces genre possèdent des profils de personnalité différents des hommes et femmes cis.[4]

En ce qui a trait aux commentaires, nous avons choisi l'attribut «body» parce qu'il contient la majorité des informations textuelles transmises par l'utilisateur. Son contenu sémantique devrait être en mesure de nous informer sur les idées, opinions et attitudes de l'utilisateur. Nous avons choisi l'attribut «wordcount» parce qu'il renseigne à la fois sur la capacité à exprimer une idée complexe, mais également sur le temps que l'utilisateur est prêt à investir sur la plate-forme Reddit. La quantité de mot rédigée permet aussi de faire une partie du traitement des données, par exemple en réalisant des traitement différents sur les commentaires en fonction de leur longueur. Nous avons choisi l'attribut «flair» parce qu'avec «body», il contient des informations et un contenu sémantique directement généré par l'utilisateur. Il permet également de mieux comprendre le contexte du commentaire, dans bien des cas. Nous avons choisi l'attribut «subreddit» parce que les communautés qu'une personne fréquente peut nous informer sur ses intérêts, ainsi que sur ses idées. Par exemple, nous nous attendons à ce que les personnes fréquentant des subreddits en lien avec les jeux vidéos ne présentent pas, en moyenne, les même score OCEAN que les personnes fréquentant des subreddits d'affiliation néo-nazie. Nous avons choisi l'attribut «score» parce qu'il indique à quel point la personne est bien intégrée dans sa niche sociale et à quel point sa contribution via le commentaire a été évaluée par ses pairs.

Pour la suite du projet, nous souhaitons également enrichir les données de différents attributs qui seront calculés automatiquement. Nous envisageons d'évaluer le sentiment associé à chaque commentaire pour au moyen d'un algorithme de détection automatique des sentiments (Section 4).Nous souhaitons également ajouter au profil la duree d'utilisation reddit en jours, et à partir de celle-ci, calculer la quantité de commentaires par jour rédigés par l'utilisateur. Une valeur standard (ex.: UNIQUE) désignera les utilisateurs n'ayant contribué qu'un seul commentaire dans le corpus de données.

Dans la table ci-après se trouvent différents attributs ainsi que la raison pour laquelle leur utilisation a été rejetée dans le cadre de notre processus prédictif.

Il est à noter que de nombreux attributs sont rejetés pour plusieurs raisons. Par exemple «state», en plus de posséder une trop grande quantité de NaN, n'est corrélé significativement avec aucune des dimensions du score OCEAN. Le corpus étant presque uniquement anglophone, il ne nous est pas pertinent de savoir combien de commentaires en anglais l'utilisateur

Données uniformes	Données éparses	Donnée redondante	Données non corrélées
enneagram	16pers_ta	mbti	is_female_proba
is_description	region	en_comments	link_id
is_percentile	state	is_native_english_country	parent_id
is_score		country	
enneagram_type		continent	
enneagram_wing		is_female	
test_scale		is_female_pred	
down		subreddit_id	

Table 2. Attributs rejetés, par raison de rejet

a écrit, ou s'il vient d'un pays anglophone.

Traitement des données

Les données contenues dans les profils et les commentaires sont particulièrement hétérogènes et possèdent une quantité importante de bruit statistique. Nous avons donc développé une méthodologie de nettoyage des données qui, nous le pensons, nous permettra plus facilement d'exploiter l'information contenues dans les données.

Comme dit précédemment, les profils et les commentaires souffrent tous deux du fléau de la dimensionalité. Énormément d'attributs ont été mesurés, et la grande majorité de ceux-ci sont soit redondants, ou non pertinents au but poursuivi. Pour résoudre ce problème, nous utiliserons une approche de diminution de la dimensionnalité. Tout d'abord une sélection manuelle des attributs les plus pertinents a été réalisée manuellement (Section 3). Nous utiliserons ensuite une procédure par sélection avant (*forward selection*) pour identifier les autres attributs qu'il convient de garder, advenant un gain suffisant en information. Un autre enjeu auquel le corpus de données, particulièrement les commentaires, nous confronte est la quantité massive d'unités statistiques. Opérer nos algorithmes sur plus de 17 millions de tuples entraîne des coûts importants en matière de calcul. Nous utiliserons donc une approche de réduction de la cardinalité. Nous avons envisagé un retrait des unités statistiques aberrantes, tels que les commentaires particulièrement longs, issus d'utilisateurs anormalement prolifiques, ou dont la quantité de token porteur de sens était en dessous d'un certain seuil. Nous avons finalement choisi une méthode de sélection orientée sur le profil psychométrique, comme il s'agit du but poursuivit. Des 17 millions de commentaires dans le corpus, nous allons donc retirer ceux dont l'auteur n'a fourni ni score OCEAN, ni MBTI. Nous utiliserons ce corpus de commentaires pour réaliser les régressions prédictives du MBTI. Par la suite, nous retirerons les commentaires des utilisateurs n'ayant pas de scores OCEAN. De cette manière, le corpus final de commentaires contiendra un peu moins de 4 millions de commentaires à traiter plutôt que 17 millions.

Notre approche générale de traitement des données sera dérivée de la famille de représentation par sac de mots (*Bag of Words*(BOW)). Nous utiliserons donc différentes techniques de prétraitement sur l'attribut «body» des commentaires pour faciliter leur représentation BOW. Les techniques choisies pour optimiser le corps des commentaire sont:

- Uniformisation de la casse
- Isolation de la ponctuation
- Lemmatisation des mots
- Normalisation des emojis en utilisant leur code UTF

- Retrait des déterminants et des conjonctions
- Remplacement des mots à trop faible cardinalité (indiquant des fautes de frappe) par un token uniforme

Ces techniques de pré-traitement ont pour but de concentrer un maximum d'information pertinentes dans un nombre le plus limité de tokens. Ce choix entraîne obligatoirement une perte d'information. Par exemple, un message entièrement écrit en majuscule (indiquant un plus haut niveau à l'échelle neuroticism) ne sera pas détecté comme tel. Cependant, nous estimons que la perte en information sera compensée par une meilleure compréhension globale du contenu sémantique par les algorithmes. En revanche, pour palier à cette perte d'information, nous créerons un nouvel attribut, pré-traitement, comptabilisant la quantité relative de majuscules utilisées dans le commentaire. D'autres adaptations, tel que le rassemblement d'emojis en familles, ou la correction automatique des mots inconnus (donc possiblement fautifs) sera considérée.

Malgré le fait que le corpus de données contienne une grande quantité d'attributs, il nous apparait que plusieurs informations pertinentes sont manquantes ou peuvent être calculées à partir des informations fournies. Notre traitement des données inclus donc une étape de bonification des données existantes, et ce sur plusieurs aspects.

Tel qu'expliqué précédemment, plusieurs attributs du profil comportent une grande quantité de valeurs manquantes. L'un de ces attributs est le score MBTI. Comme nous pensons qu'il s'agit d'un prédicteur majeur du score OCEAN de l'utilisateur, nous allons créer un algorithme permettant de l'inférer. Après quelques analyses exploratoires, nous pensons que la meilleure manière de générer une valeur assez proche de la valeur réelle pour permettre la prédiction des scores OCEAN est la régression linéaire à partir des statistiques des commentaires de l'utilisateur. Nous avons constaté que différentes caractéristiques (ex. l'extraversion) possèdent une signature spécifique en terme de mots utilisés plus que la moyenne, ainsi qu'une longueur moyenne différente. Notre but est que la régression infère individuellement chacun des 4 traits du MBTI en comparant les messages de l'utilisateur (représenté sous forme de vecteur BoW) à la signature du trait et de son opposé (ex. introversion). L'algorithme sera alors en mesure de déterminer de laquelle de ces deux signatures le vecteur du message se rapproche le plus. Ces valeurs seront accumulées de message en message (en utilisant la méthode du *gradient ascend*) pour arriver à un score final. De cette manière, les utilisateurs ayant de nombreux messages n'accumuleront pas nécessairement beaucoup plus de points en raison de la quantité de message. Nous envisageons également de mettre une limite arbitraire au nombre maximal de message pris en compte dans la sommation pour limiter l'impact de la disparité du nombre de messages dans le corpus. Pour les utilisateurs possédant un grand nombre de messages, il pourra être envisagé de faire un partitionnement et réaliser une moyenne des scores obtenus.

Puisque plusieurs échelles du score OCEAN sont en lien avec l'émotivité de la personne (ex. neuroticisme), nous pensons qu'il est important pour notre algorithme prédictif d'être en mesure de comprendre l'information émotive contenue dans les commentaires de l'utilisateur. Notre analyse des sentiments sera réalisée à l'aide de l'algorithme Roberta Base Go Emotions, disponible publiquement. Cet algorithme permet d'attribuer à une chaîne de caractères un score numérique allant de 0 à 1 pour 28 émotions différentes. Nous l'avons choisi car il a été entraîné sur le GoDataset, un corpus de données de 57000 commentaires Reddit. Le corpus 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. Ci-dessous se trouve un

exemple d'analyse réalisée sur les commentaires d'un seul utilisateur. Chaque point de données représente la moyenne de 50 commentaires consécutifs en ordre chronologique (Fig.4). Ce graphique permet de suivre adéquatement l'évalution de l'état d'esprit de l'utilisateur à travers le temps. Une analyse par série temporelle (*time series*) permettrait de constater les motifs de variation dans les émotions exprimées par l'utilisateur au fil du temps, ainsi que les périodes émotionnellement hors de l'ordinaire. Nous pensons donc que ces données permettront d'aider à peaufiner la prédiction du score OCEAN.

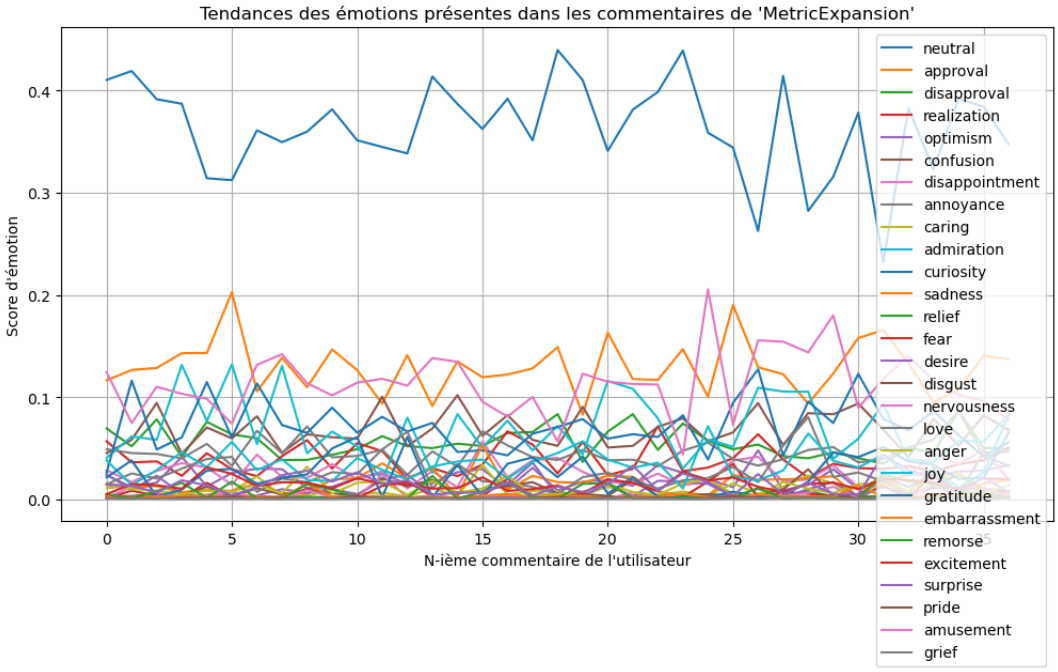

Figure 4. Tendances des émotions de l'utilisateur «MetricExpansion»

L'intensité de l'activité de l'utilisateur sur la plateforme est, en elle-même, une information révélatrice de certains traits de personnalité. Elle ne se trouve cependant pas dans le profil, et nous devrons donc l'inférer. Pour ce faire, nous utiliserons deux attributs: duree_reddit et msg_par_jour. La durée reddit sera calculée en identifiant le premier et le dernier commentaire posté sur la plateforme dans le jeu de donnée au moyen de l'horodatage de chaque commentaire. La différence entre ces deux points sera ensuite calculée et constituera la durée d'activité sur la plateforme. La quantité de message par jour sera calculée par une division de l'attribut num_message par duree_reddit. Les utilisateurs n'ayant posté qu'un commentaire recevront la valeur correspondant à une journée (24h).

Une fois le pré-traitement et la bonification des données réalisées, nous planifions d'utiliser une approche BoW. Les vecteurs BOW permetront d'analyser les tendances d'utilisation des mots selon les scores OCEAN. Au cours de notre analyse, nous avons prototypé une méthode identifiant les mots les plus fortement corrélés à différentes échelles OCEAN à l'aide d'un modèle de regression linéaire multivarié. Notre prototype analyse les fréquences d'utilisation de chaque mot dans un vocabulaire(élevé à la racine carrée) et les scores OCEAN. Ce prototype nous as permis d'identifier différentes limites à cette approche. Les corrélations les

plus fortes restent assez faibles et ne permettent pas, utilisées seules, de faire une prédiction exacte du score OCEAN. La deuxième limite vient de la présence de nombreuses données aberrantes. Par exemple, plusieurs utilisateurs tendent à utiliser certains mots de manière disproportionnée. De plus, puisque les proportions d'utilisation de chaque mot sont généralement faibles (<0.0025 dans notre prototype), une personne répétant un mot ad nauseam aura un effet important sur la distribution des scores. Pour contrer ces limites, nous utiliserons plutôt une proportion de mots utilisés plutôt que la quantité de fois où ils sont utilisés. Également, puisque leur corrélation au score OCEAN est faible, nous pairerons cette analyse avec d'autre méhodes, comme les N-grammes et l'analyse de sentiment pour augmenter les chances de réaliser une prédiction exacte.

Une partie significative des utilisateurs pour qui le score OCEAN a été omis possèdent un nombre de commentaire extrêmement limité (<50). Certains d'entre-eux ne possèdent qu'un ou deux commentaires. Pour cette raison, nous pensons qu'inférer le score OCEAN de ces utilisateurs à l'aide de l'approche BoW ne sera pas possible. Leurs données seront soumises à une forme alternative de traitement. Pour le moment, l'approche envisagée est de prédire que ces gens possèdent le score médian pour chaque sous-échelle du Big Five. Nous sommes conscient que cette approche est statistiquement improbable. En toute logique, la population de personnes qui ne postent que quelques messages sur Reddit possède probablement des caractéristiques qui lui sont propres et expliquent en partie leur faible volume de contribution à la plateforme. Il peut par exemple s'agir de comptes «*throwaway*», à partir duquel un utilisateur poste des commentaires qu'il ne souhaite pas voir relié à son nom d'utilisateur principal. De par leur nature, ces comptes sont plus prompt à partager du contenu sensible, ou tenir des propos controversés. Il peut également s'agir d'utilisateurs qui utilisent généralement Reddit en lecture seulement et qui, pour une raison inconnue, se sont sentis interpelés par une discussion spécifique. Ces deux populations possèderaient des attributs OCEAN différents de la médiane, mais en l'absence d'indication sur la nature du compte, l'application du score median nous semble statistiquement l'approche la plus à même de minimiser l'écart entre le score réel et le score prédit pour ces utilisateurs. Une partie restreinte d'utilisateurs pour qui le score OCEAN a été omis possèdent un nombre de commentaire significativement au dessus de la moyenne. De par l'abondance d'information contenues dans les commentaires, nous allons implémenter dans l'algorithme prédictif un calcul pondéré du score. Plus un utilisateur possède de messages, plus l'algorithme accordera de poids à la prédiction par commentaires, et moins à la prédiction à l'aide du MBTI et autres caractéristiques du profil. À l'inverse, moins un utilisateur aura rédigé de commentaires, plus le poids du score prédit à l'aide des attributs du profil sera élevé, avec un poids plus faible accordé au score issu des commentaires.

Le profil des utilisateurs possède beaucoup de champs, mais la plupart des attributs contenus prennent un nombre relativement limité de valeur, ou alors sont de nature nominale ordonnée. Pour faciliter l'utilisation des algorithmes et éviter que certains attributs possédant de nombreuses valeurs ne prenne l'ascendant sur d'autres attributs, nous allons transformer les catégories soit en valeurs binaires (ex. «genre» devient «is_woman» et «is_trans») ou en valeurs sur un intervalle [0,1] (ex. l'âge).

Architecture de prédiction

Comme dit précédemment, les utilisateurs peuvent être divisés en différents groupes nécessitant des modèles différents pour obtenir une prédiction précise en fonction de la quantité

de données disponibles sur eux. La quantité de commentaires, notamment, influe beaucoup sur le type de prédiction qu'il nous est possible de réaliser. Deux options s'offrent à nous:

(1) Entraîner un modèle se basant uniquement sur les données du profil, et un second prenant également en compte les commentaires. En dessous d'un nombre arbitraire de commentaires, nous utilisons le premier modèle. Les utilisateurs restants sont prédits à l'aide du second modèle.
(2) Implémenter le modèle expliqué précédemment utilisant une approche pondérée du score en fonction de la quantité de commentaires de l'utilisateur.

Nous espérons produire un modèle réunissant les deux corpus de données. Nous pensons que notre meilleure chance d'y arriver est d'unifier nos données en extrayant les mots utilisés par chaque utilisateur et en identifiant les corrélations entre ces mots et leurs scores OCEAN. Le rapport $\frac{\text{Mots de la liste utilisés par l'utilisateur}}{\text{Cardinalité de la liste}}$ sera introduit comme attribut dans le profil. Cette approche privilégie l'interprétabilité, garantissant que cette partie de la prédictions du modèle puisse être facilement comprises et analysées. Elle est cependant biaisée en faveur des utilisateurs postant fréquemment, et rédigeant de longs commentaires.

Exemples tirés de la liste de la sous-échelle Openness: [*'dude', 'point', 'hour', 'month', 'alita', 'high', 'rather', 'giveaway', 'several', 'important', 'btw', 'reason', 'cute', 'fair', 'couple', 'schizoid', 'edit', 'whenever', 'link', 'believe', 'intp', 'ship', 'tho', 'nice', 'give', 'nation', 'military', 'single', 'feminist',...*]

En tandem avec l'étape décrite précédemment, nous allons décomposer les commentaires sous forme de N-grammes. Cette stratégie implique un coût plus important en ressources, mais risque d'arriver à des résultats plus intéressants, et tout aussi interprétables.

La sémantique extraite lors de ces deux étapes concurrentes sera combinée à l'analyse de sentiments (obtenue par apprentissage de réseaux de neurones). L'information résultante sera ensuite encodée dans un format compatible avec le modèle prédictif sur les données du profil:

- Des statistiques descriptives en lien avec le score émotionnel, comme la moyenne et la variance pour certaines émotions-clés, ou le rang de chaque émotion en terme de présence dans les messages. Nous postulons qu'une personne avec un plus haut score à la sous-échelle « *neuroticism* » aura une haute moyenne d'angoisse et de colère, par exemple.
- Une prédiction directe du score OCEAN par un réseau de neurones récurrents(RNN) sur une représentation de l'information sémantique et émotive (poteniellement lissée) au fil du temps. Cette prédiction servira à un autre modèle qui l'unira avec les données du profil.

La représentation graphique de l'architecture de notre algorithme prédictif est la suivante (Fig.5):
Nous envisageons également une étape supplémentaire à entre l'utilisation de l'algorithme #1 et la prédiction finale. Celle-ci est cependant encore en exploration et n'a donc pas été ajoutée au modèle graphique. Nous pensons qu'utiliser un modèle d'ajustement de cohérence (MAC) pourrait se révéler utile. Malgré le fait que la littérature scientifique en psychométrie déclare que les sous-échelles OCEAN sont orthogonales (i.e. entièrement indépendantes les unes des autres)[5], plusieurs études nous permettent de douter que ce soit réellement le cas [6][7]. Advenant le cas où certaines sous-échelles étaient corrélées, un MAC permettrait, à l'aide d'une fonction de densité, d'ajuster les scores prédits pour maximiser la probabilité du point trouvé. Les performances de ce modèle restent à tester avant d'envisager de l'implémenter dans notre algorithme final.

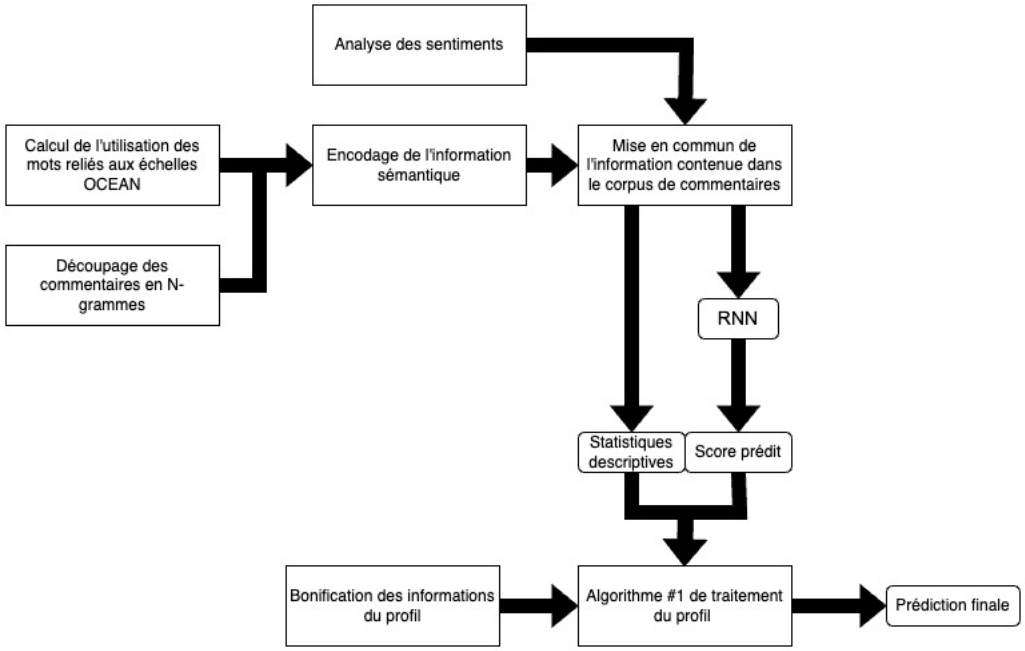

Figure 5. Architecture prévue de l'algorithme prédictif

Note sur l'utilisation de données supplémentaires: Le corpus ne contient que 1400 scores OCEAN étiquetés. Nous envisageons de réaliser une recherche des banques de données disponibles publiquement pour tenter de trouver plus d'instance où les scores OCEAN de différentes population ont été mesurés. Nous pensons qu'un apport en données faciliterait l'entrainement de notre algorithme prédictif. Par contre, si nous trouvons une autre source de données, nous devrons tenir compte du fait que celles-ci ne sont pas normalisées de la même manière que pour le corpus PANDORA. Ce processus de normalisation devrait être répliqué par nos soins. De plus, peu importe la population sur laquelle les nouveaux scores OCEAN ont été collectés, il est probable que ses caractéristiques soient intrinsèquement différentes de celles de la population de Reddit en 2015. Entraîner un modèle sur des populations différentes nécessiterait une évaluation des distributions de chacune des sources.

Procédure de test

Pendant le développement de l'algorithme, les métriques utilisés pour mesurer l'apprentissage et le niveau de qualité de la prédiction sera majoritairement la différence moyenne entre la prédiction et le score réel du même utilisateur (*Root Mean Square Error* (RSME)). Durant cette étape, aucun entraînement ne sera réalisé avec des données dont la valeur réelle est inconnue. Un coefficient de variation (R^2) sera également calculé pour mieux comprendre la proportion de variation attribuable à différents attributs dans notre modèle. D'autres métriques pourront être explorés en cours de développement, comme par exemple une matrice de confusion à seuil d'acceptation pour comprendre la performance de notre régression linéaire des échelles MBTI.

Nous utiliserons une méthode de validation finale par retenue des données (*holdout method*). Cette méthode est sélectionnée en raison de la quantité relativement faible d'unités statistiques disponibles à la création de partitions d'entrainement (1408 unités statistiques

possédant un score OCEAN complet connu). Nous souhaitons mettre de côté un échantillon de validation équivalent à 10% des unités statistique, sur lequel les algorithmes prédictifs n'auront pas été entrainés. Par la suite, une fois l'algorithme entrainé, les données de validation seront utilisées pour calculer la précision de notre modèle en comparant la prédiction avec la donnée de validation.

D'autres méthodes de test, notamment la validation croisée k-échantillons (*k-fold cross-validation*), ont également été envisagées, mais n'ont finalement pas été retenues. La principale raison pour laquelle cette méthode a été rejetée est que le partitionnement des données risque de créer des échantillons de cardinalité inférieure à 100, ce qui correspond à l'étendue des scores des échelles OCEAN. De ce fait, il est probable que certains scores ne soient pas représentés dans chaque échantillon, ce qui pourrait biaiser les résultats, voire fausser les prédictions d'un algorithme n'ayant vu aucune donnée associée à une plage de scores particulière.

Mentions

L'équipe de travail déclare ne pas être en situation de conflit d'intérêt. 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] 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.

[2] 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/.

[3] J. Hartung, M. Bader, M. Moshagen, and O. Wilhelm. "Age and gender differences in socially aversive ("dark") personality traits". In: *European Journal of Personality* 36.1 (2022), pp. 3–23. DOI: 10.1177/0890207020988435. eprint: https://doi.org/10.1177/0890207020988435. URL: https://doi.org/10.1177/0890207020988435.

[4] C. Rodriguez-Seijas, T. A. Morgan, and M. Zimmerman. "Transgender and Gender Diverse Patients Are Diagnosed with Borderline Personality Disorder More Frequently Than Cisgender Patients Regardless of Personality Pathology". In: *Transgender Health* 9.6 (2024), pp. 554–565. DOI: 10.1089/trgh.2023.0062. eprint: https://doi.org/10.1089/trgh.2023.0062. URL: https://doi.org/10.1089/trgh.2023.0062.

[5] H. Eysenck. "Four ways five factors are not basic". In: *Personality and Individual Differences* 13.6 (1992), pp. 667–673. ISSN: 0191-8869. DOI: https://doi.org/10.1016/0191-8869(92)90237-J. URL: https://www.sciencedirect.com/science/article/pii/019188699290237J.

[6] G. J. Boyle, L. Stankov, and R. Cattell. "Measurement and Statistical Models in the Study of Personality and Intelligence". In: *Humanities Social Sciences papers* (Jan. 1995). DOI: 10.1007/978-1-4757-5571-8_20.

[7] J. Musek. "A general factor of personality: Evidence for the Big One in the five-factor model". In: *Journal of Research in Personality* 41.6 (2007), pp. 1213–1233. ISSN: 0092-6566. DOI: https://doi.org/10.1016/j.jrp.2007.02.003. URL: https://www.sciencedirect.com/science/article/pii/S0092656607000256.