Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Étapes préparatoires pour la détection des valeurs humaines dans des commentaires du domaine de la parfumerie

Boyu NIU[1, 2]

(1) ERTIM de l'Inalco, 2 rue de Lille, 75007 Paris, France
(2) IFF, 61 rue de Villiers, 92200 Neuilly-sur-Seine, France
boyu.niu@{inalco.fr, iff.com}

RÉSUMÉ

La détection des valeurs humaines dans le texte est une tâche qui intéresse les industriels dans la mesure où elles complètent le profil des consommateurs. Cette détection nécessite des outils et des méthodes issues du traitement automatique des langues (TAL) et s'appuie sur un modèle psychologique. Il n'existe que très peu de travaux, alliant modèles psychologiques de valeurs humaines et extraction de leur réalisation linguistique sur les réseaux sociaux à l'aide du TAL. Dans cet article, après avoir défini le modèle des valeurs de Schwartz que nous utilisons ainsi que le corpus en cours de construction pour le domaine de la parfumerie, nous proposons quelques pistes de réflexion possibles pour la construction de technologies permettant de relier des marqueurs textuels à des valeurs humaines.

ABSTRACT

Detecting Human Values in Comments on Perfumery

The detection of human values from texts is an interesting work for enterprises as this is a way to create a more comprehensive profile of consumers. This task requires tools and methods of automatic language processing (NLP) and relies on a psychological model. There are very few works that combine psychological models of human values and the extraction of their linguistic realization on social networks using NLP. In this article, after having defined the model of human values of Schwartz that we use, and the corpus being collected for the field of perfumery, we propose some possible ideas for the implementation of technologies to find links between textual signals and human values.

MOTS-CLÉS : Modèle des valeurs humaines, détection automatique, TAL, études psycholinguistiques, parfumerie

KEYWORDS: Human Value Models, Automatic Detection, NLP, Psycholinguistic studies, perfumery

1. Introduction

Les commentaires sur les produits publiés sur Internet constituent une source précieuse pour le monde industriel. D'un côté, ils permettent aux fabricants d'avoir un retour rapide sur comment leurs produits sont perçus chez les consommateurs et d'améliorer ainsi leurs offres sur le marché. D'un autre côté, comme chacun a sa propre manière de s'exprimer, ces mêmes commentaires peuvent aider les fabricants à mieux connaître les profils des utilisateurs de leurs produits. C'est particulièrement le cas des produits qui ont un caractère émotionnel, comme les parfums et autres produits parfumés, tels que les shampoings, les gels douche, les lessives ou encore les parfums d'ambiance.

Notre travail a pour objectif de détecter les différents profils de consommateurs pour les parfums et les produits parfumés, en utilisant des méthodes du traitement automatique des langues (TAL), et en s'appuyant sur un modèle de valeurs humaines proposé par Schwartz (1992, 1996, 2003, 2006). Dans cet article, après avoir rapidement présenté ce modèle de valeurs, nous nous attardons sur les méthodes TAL pouvant servir à atteindre notre objectif. Enfin, nous présentons les problématiques liées à la constitution du corpus utile pour mener à bien ce projet, avant de conclure sur ce que sont les prochaines étapes de notre travail.

2. Le système des valeurs humaines de Schwartz

Les valeurs humaines sont étudiées et utilisées dans le monde de la psychologie. La notion de valeur qui nous intéresse est définie dans le dictionnaire Larousse comme étant « ce qui est posé comme vrai, beau, bien, d'un point de vue personnel ou selon les critères d'une société et qui est donné comme un idéal à atteindre, comme quelque chose à défendre. » Les valeurs peuvent ainsi expliquer les choix que font les gens dans leur vie (Verplanken et Holland, 2002). Pour les entreprises, les valeurs associées à chaque consommateur peuvent aider à comprendre leur choix de produits. Ce travail s'inscrit dans cette hypothèse et cherche à détecter

dans les commentaires mêmes des consommateurs les raisons sous-jacentes à leur choix de produits et de les relier, dans la mesure du possible, à un ensemble de valeurs auxquelles ils adhèrent, qui les caractérisent.

Le modèle proposé par Schwartz contient les dix valeurs de base :
Autonomie (*Self-Direction*) est définie par la pensée et l'action indépendantes : choisir, créer et explorer.
Stimulation : une valeur motivée par l'excitation, la nouveauté et les défis dans la vie.
Hédonisme (*Hedonism*) : une valeur caractérisée par le plaisir personnel, la gratification sensuelle, le fait de vouloir profiter de la vie.
Réussite (*Achievement*) se définit par quelqu'un qui démontre ses compétences et qui veut obtenir de l'approbation sociale.
Pouvoir (*Power*) est défini par la recherche du statut social et du prestige, et le contrôle et la domination sur les personnes et les ressources.
Sécurité (*Security*) est caractérisée par la sécurité, l'harmonie, et la stabilité de la société, des relations et de soi-même.
Conformité (*conformity*) est une valeur qui insiste sur la maîtrise de soi dans les interactions quotidiennes, surtout avec les proches (obéissant, politesse, honorer ses parents et les personnes âgées).
Tradition est caractérisée par le respect pour la tradition et le fait d'être humble.
Bienveillance (*Benevolence*) met l'accent sur le bien-être des proches (la vraie amitié, l'amour mûr).
Universalisme (*Universalism*) est défini par la compréhension, l'appréciation, la tolérance et la protection du bien-être de tous les êtres humains et de la nature.

Pour mesurer les valeurs d'un individu, Schwartz avait proposé le questionnaire SVS (*Schwartz Value Questionnaire*) (Schwartz, 1992) qui comporte 57 questions composées essentiellement de noms et d'adjectifs pour lequel les répondants sont invités à donner un score selon l'importance que représente la notion portée par ces mots dans leur vie. Plus tard, Schwartz et al. (2001) ont mis au point un autre questionnaire PVQ (*Portrait Value Questionnaire*), qui contient 40 « portraits », c'est-à-dire des descriptions d'un individu imaginé, où les répondants sont invités à dire à quel degré cet individu leur ressemble. Schwartz (2003) a également publié une version plus courte du PVQ qui contient 21 questions.

3. Méthodes pour la détection des traits psychologiques

Les valeurs humaines peuvent être considérées comme des caractéristiques psychologiques d'un individu. Les traits de personnalité sont un autre système qui est souvent utilisé pour caractériser les individus : alors qu'à notre connaissance peu de travaux ont été faits pour détecter les valeurs de Schwartz dans les textes, différentes expérimentations ont été menées pour la détection des traits de personnalité en s'appuyant sur des méthodes de TAL. Nous pouvons donc nous inspirer de ces méthodes dans notre projet de recherche. Dans cette partie, nous présentons brièvement les deux modèles de personnalité les plus utilisés dans la littérature, avant de décrire les liens entre le modèle Big-5 et celui des valeurs humaines trouvés par d'autres chercheurs. Ensuite nous nous plongeons dans les méthodes existantes pour la détection de ces traits de personnalité en utilisant des méthodes de TAL. Enfin, nous présenterons les quelques publications que nous avons pu identifier en lien avec la détection des valeurs humaines à partir du texte.

3.1. Big-5 et MBTI

Big-5 et MBTI sont deux modèles qui sont beaucoup utilisés dans le domaine des études des personnalités. Le modèle de Big-5 (Digman, 1990) est aussi connu sous le nom du modèle « OCEAN » en raison du nom des cinq dimensions qu'il mesure :
Ouverture : un individu qui a de la curiosité et qui aime les nouvelles choses versus un individu dogmatique et prudent ;
Conscience : un individu organisé et efficace versus un individu insouciant ;
Extraversion : un individu extraverti qui aime parler avec d'autres gens versus un individu réservé ;
Amabilité : un individu franc, généreux, humble et digne de confiance versus un individu compliqué, soupçonneux et antagonique envers les autres ;
Neuroticisme : un individu sensible et nerveux versus un individu qui a de la confiance en soi.
Il existe un questionnaire « *Big-5 Inventory* » (John et Srivastava, 1999) qui permet de donner un score à chacune de ces dimensions d'un individu.

Le modèle MBTI (Briggs Myers et al., 1998), quant à lui, mesure les quatre dimensions suivantes :
Orientation de l'énergie : **Extraversion / Introversion** ;
Recueil d'information : **Sensation / Intuition** ;
Prise de décision : **Pensée / Sentiment** ;
Mode d'action : **Jugement / Perception**.

Nous pouvons constater que certaines dimensions mesurées dans ces deux modèles ont des points communs avec le modèle de valeurs de Schwartz.

3.2. Liens entre les traits de personnalité et les valeurs humaines

Les traits de personnalité font référence à ce que sont les individus, tandis que les valeurs décrivent ce qu'ils considèrent comme important. Roccas et al. (2002) montrent que les valeurs et les traits de personnalités sont des constructions psychologiques conceptuellement et empiriquement distinctes, mais elles sont liées. Les expériences menées respectivement par Luk et Bond (1993) et par Roccas et al. (2002) indiquent qu'il existe des corrélations entre certains traits de personnalité et certaines valeurs. Par exemple, le trait « amabilité » a une corrélation positive avec les valeurs « bienveillance » et « tradition », ainsi qu'une corrélation négative avec les valeurs « pouvoir » et « réussite ».

3.3. Corpus existants pour la détection des traits de personnalité

Il existe différents corpus textuels pour l'étude des traits de personnalité dans la littérature.
Pennebaker et King (1999) ont, pendant plusieurs années, demandé à des étudiants en psychologie d'écrire des essais et les ont invités à répondre au questionnaire du Big-5. Ils ont ainsi créé un corpus qui contient à la fois la production textuelle de ces étudiants et les traits de personnalité de chacun d'eux. Cela nous permet de faire une étude comparative entre le texte et la personnalité de son auteur.
Le corpus de Pennebaker et King a été créé sans interaction : c'est-à-dire que les étudiants ont eu pour consignes d'écrire des essais, et il n'y avait pas de modérateurs entretenant des conversations avec eux. PerSIA (Ivanov et al., 2011), quant à lui, est un corpus composé des dialogues téléphoniques transcrits en textes (en italien). Les participants répondent au questionnaire du Big-5, jouent le rôle d'un touriste ou celui d'un agent d'un centre d'information du tourisme. Leurs conversations sont enregistrées avant d'être transcrites.
Un autre corpus bien connu est « MyPersonality » (Stillwell et Kosinski, 2011), également basé sur le modèle Big-5. C'était une application Facebook qui permettait à des utilisateurs de participer à des recherches psychologiques en remplissant un questionnaire de personnalité, et les posts publiés par ces utilisateurs sont inclus dans ce corpus. Il n'est plus accessible au public depuis 2018.
D'autres corpus sont basés sur le modèle MBTI. Par exemple, le corpus MBTI Kaggle [1] contient des posts publiés par 8600 internautes sur un forum de personnalité, ainsi que leurs traits de personnalité déclarés par eux-mêmes.
MBTI-Twitter (Plank et Hovy, 2015) est un corpus qui contient les tweets ainsi que les traits de MBTI auto-déclarés par leurs auteurs.

3.4. Détection des traits de personnalité avec le TAL

3.4.1. Lexiques psycholinguistiques et méthodes statistiques

Pennebaker et King (1999) ont trouvé un lien entre le style d'écriture (usage de certaines catégories grammaticales de mots, par exemple) d'un individu et ses traits de personnalité. Ils ont classé les mots en plus de 80 catégories psycholinguistiques en fonction de leur partie du discours, leurs valeurs sémantiques, leur longueur, etc., et ont proposé un système de prédiction de personnalité LIWC (*Linguistic Inquiry and Word Count*) en s'appuyant sur ces classes de mots et des méthodes statistiques (Pennebaker et al., 2001, 2015).
Aux traits de LIWC, Mairesse et al. (2007) ont ajouté un autre système de traits psycholinguistiques MRC (Coltheart, 1981) pour prédire les traits de personnalité d'un individu à partir de sa production textuelle. Le système MRC contient plus de 150 000 mots, auxquels sont associées des informations comme l'estimation de l'âge d'acquisition, la fréquence et la familiarité. Ensuite, les auteurs ont testé des méthodes de classification (arbre de décision, SVM, etc.), de régression (linéaire, M5, etc.) et de modèle de ranking qui ont donné tous les trois des résultats meilleurs que la *baseline* de la classe majoritaire.
En plus de LIWC et MRC, Poria et al. (2013a) utilisent SenticNet 2.0 (Cambria et al., 2012) pour calculer la polarité du sentiment exprimé dans un texte donné, EmoSenticNet (Poria et al., 2013b) pour la détection des émotions, ainsi que EmoSenticSpace (Havasi et al., 2009) et ConceptNet (Havasi et al., 2007) pour intégrer des connaissances générales sur le monde (*common sense knowledge*) dans le système. Ces traits-là sont ensuite passés à un classifieur pour la prédiction de la personnalité.
Park et al. (2014) n'ont pas utilisé de lexique linguistique prédéfini. En revanche, ils créent des traits linguistiques à partir du corpus étudié, i.e. les mots et les expressions bigrammes / trigrammes fréquentes présents dans le corpus, ainsi que les thématiques identifiées à l'aide de LDA (Blei et al., 2003). Ensuite une réduction de dimensions est faite avant la régression pour la prédiction de la personnalité.

[1] https://www.kaggle.com/datasnaek/mbti-type

3.4.2. Plongement des mots + réseaux de neurones

Majumder et al. (2017) ont utilisé le Word2Vec (Mikolov et al., 2013) pour convertir les mots en vecteurs, et ont passé ces vecteurs à un réseau de neurones convolutifs (CNN), avant de faire la prédiction sur chacune des dimensions du modèle Big-5 avec un classifier basé sur machine à vecteurs de support (SVM) ou perceptron multicouche (MLP).
En raison de la séquentialité des mots dans le texte, Hernandez et Ian Scott (2017) ont fait des expériences avec des modèles de réseaux de neurones récurrents (RNN), y compris le RNN simple, GRU, LSTM et Bi-LSTM. Ils rapportent que le LSTM donne le meilleur résultat.
Darliansyah et al. (2019) ont vectorisé les mots avec GloVe (Pennington et al., 2014) et ont fait des expériences avec à la fois le CNN et la LSTM. Ce qui est le plus différent des deux études citées précédemment est que la polarité du sentiment exprimé dans le texte y est prise en considération de manière explicite.

3.4.3. Transformers

Il y a récemment des études sur la détection des traits de personnalité avec le modèle BERT (Devlin et al., 2018). Keh et Cheng (2019) montrent que l'affinage du BERT donne un meilleur résultat de prédiction par rapport à SVM, LSTM et régression logistique sur un corpus MBTI. Arijanto et al. (2021) montrent que le BERT donne un meilleur résultat de prédiction quand on affine les paramètres du BERT et l'utilise en tant que classifier sur un corpus Big-5 composés des tweets.
Leonardi et al. (2020) utilisent le BERT comme encodeur pour vectoriser les phrases, avant de passer ces vecteurs à un réseau de neurones pour prédire le score d'une certaine dimension du modèle Big-5.

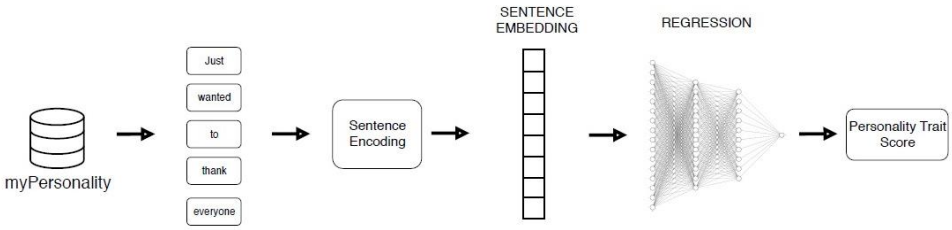

FIGURE 1. Le pipeline proposé par Leonardi et al. (2020)

Comme Leonardi et al. (2020), Ren et al. (2021) utilisent aussi le BERT comme encodeur. Et comme Darliansyah et al. (2019) cités précédemment, ils tiennent aussi explicitement compte de la polarité du sentiment exprimé dans le texte. À la fin, ils utilisent un réseau de neurones (CNN ou RNN) et ensuite prédisent un résultat.

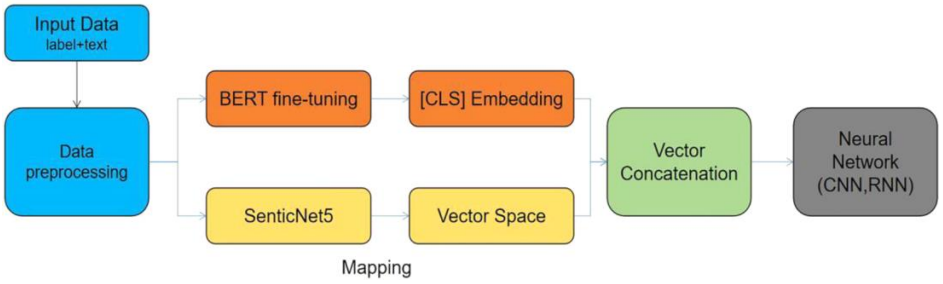

FIGURE 2. Le pipeline proposé par Ren et al. (2021)

En plus du modèle BERT de base, Vásquez et Ochoa-Luna (2021) ont testé des variants du BERT comme RoBERTa (Liu et al., 2019) et un autre modèle de la famille des Transformers, le XLNet (Yang et al., 2019) sur un corpus MBTI. Ils rapportent que les meilleurs résultats sont obtenus avec les modèles DistilBERT (version de base) et RoBERTa (version distillée).

3.5. Faisabilité de la détection de la personnalité à partir des tweets

Si Pennebaker et King. (1999), ainsi que Tausczik et Pennebaker (2010) ont trouvé des liens entre le langage et la personnalité de manière générale, il est sans doute judicieux de se demander s'il est possible de détecter des traits de personnalité, ou d'autres traits psychologiques, dans un certain type de texte. Par exemple, Stajner et Yenikent (2021) ont étudié les liens éventuels entre les tweets et la personnalité de leur auteur. Ils ont mis en avant l'idée qu'il n'existerait pas de « traduction » équivalente entre le résultat du questionnaire MBTI d'un individu et ses tweets après une expérimentation (sauf la dimension Extraversion / Introversion). Celle-ci leur donne des arguments pour les hypothèses suivantes : les tweets ne contiennent pas suffisamment de signaux pour la détection des traits MBTI, et les données textuelles ne résonnent pas bien avec les résultats du questionnaire MBTI (d'après les auteurs,

l'approche utilisée pour construire le modèle MBTI a eu un nombre considérable de critiques, car ce n'est pas une approche fondée sur les données).

3.6. Travaux existants pour la détection des valeurs humaines à partir du texte

Les travaux directement liés à la détection des valeurs humaines que nous avons pu trouver sont beaucoup moins nombreux que ceux liés à la détection des traits de personnalité dans le domaine du TAL. Takayama et al. (2014) utilise un autre modèle de valeur, celui de valeurs de Cheng et Fleischmann (2010), et proposent un modèle de prédiction basée sur la variable latente probabiliste (*Latent Value Model*) au niveau des mots pour détecter des valeurs dans des phrases. Plus tard, la même équipe a ré-utilisé ce modèle pour analyser des articles de presse après le séisme de Fukushima au Japon (Ishita et al., 2017). Li et al. (2019) ont montré un système d'analyse textuelle qui contient une fonctionnalité de la détection des valeurs (du système de Schwartz) de la population d'une région, mais ils ne précisent pas les méthodes utilisées.

4. Objectifs futurs

Comme on l'a vu, le système des valeurs humaines et celui des traits de personnalité sont tous les deux des outils qui peuvent être utilisés pour caractériser les individus d'un point de vue psychologique. Nous pouvons ainsi faire des expériences avec les méthodes de détection de la personnalité appliquées à la détection des valeurs de Schwartz, et observer si nous obtenons les mêmes niveaux de performance. Notre domaine étant celui de la parfumerie il est nécessaire de constituer un corpus particulier.

4.1. Constitution du corpus

4.1.1. Questionnaire et enquête

Pour constituer notre corpus, nous nous sommes inspirés des corpus cités dans le paragraphe 3.2 et avons fait appel à une société d'études consommateurs pour organiser cette enquête. 2000 personnes âgées de 16 ou plus en France et aux États-Unis ont respectivement été invitées à participer à notre enquête en répondant à un questionnaire en ligne.

Ce questionnaire contient trois parties principales. La première partie est composée des questions sur les valeurs, c'est-à-dire le PVQ-21 que nous avons cité dans le paragraphe 2 de cet article, avec de légères modifications. La deuxième partie du questionnaire contient quelques questions sur les informations démographiques, comme l'âge et le genre des répondants. Dans la troisième partie, les participants sont d'abord invités à choisir les grandes catégories de produits qu'ils utilisent (parfums, produits de soin du corps, produits d'entretien de la maison, fragrance d'ambiance, et produits d'entretien des vêtements). Ensuite, ils peuvent renseigner la marque et le nom du produit qu'ils utilisent pour chacune des catégories choisies (ils peuvent soit taper des caractères, soit prendre simplement une photo du produit). Enfin et surtout c'est ici qu'on les invite à écrire un commentaire sur ce produit comme ils le feraient sur un réseau social. Ils doivent écrire au moins 200 caractères pour valider leur commentaire. Ces enquêtes ont été réalisées en anglais et en français.

4.1.2. Un premier échantillon du corpus

L'enquête est en cours, mais nous avons déjà obtenu un premier échantillon du corpus.

L'échantillon américain contient 497 commentaires écrits par 116 participants. Le nombre moyen de tokens par commentaire est de 43,81, et le nombre moyen de caractères par commentaires est de 238,24.
Voici quelques-uns de ces commentaires :

I really like this perfume. It smells really good and is reasonably priced. It is a really good product for the cost of what it is. I have been happy with all the perfumes I have gotten from Juicy Couture. I am completely satisfied with them.

Cleans my skin with just the right amount of fragrance. The bar is a convenient size and lasts for a reasonable time. The bar really lathers up better than some other brands I have tried. This bar is a great price.

It provides a very pleasant scent. It removes odors from our household well. This is a positive brand that provides us with what we need to accomplish to accomplish our goals. We find the company to be very responsible.

Dès à présent nous voyons clairement apparaître, dans les exemples précédents, un ensemble de thématiques ou de signaux liés à la motivation d'achat.

Ainsi :

- "I really like this perfume. It smells really good" et "It provides a very pleasant scent." font référence à l'appréciation du parfum ;
- "is reasonably priced. It is a really good product for the cost of what it is" et "This bar is a great price" font référence au concept de prix ;
- "lasts for a reasonable time" fait référence au concept de la tenue du parfum ;
- "the right amount of fragrance" fait référence au concept d'intensité du parfum ;
- "Cleans my skin" et "It removes odors from our household well" font référence aux effets du produit.

L'échantillon français contient 397 commentaires écrits par 94 participants. Le nombre moyen de tokens par commentaire est de 37,77, et le nombre moyen de caractères par commentaire est de 238,82.
Voici quelques-uns de ces commentaires :

Parfum intense qui laisse un très bon sillage. Il tient toute la journée. Il est parfait aussi pour une soirée. Il sent très bon je l adore. Convient plutôt à une femme ou une jeune femme qu à une adolescente.

Produit efficace qui respecte la nature et remplit ses objectifs. Issu de techniques respectueuses de l'environnement et produit par une société locale. Je ne peux que recommander ce produit. Le prix est un peu cher

bougie d'ambiance assez insatisfaisante compte tenu de sa composition generant une pollution de l'intérieur du ligement sinon belle lumière et flamme tres stable et assez jolie dommage pour la composition

Dans le corpus français on retrouve des thématiques qui sont les mêmes que celles du corpus américain.
Ainsi :

- « Il sent très bon je l adore » est un marqueur de l'appréciation du parfum ;
- « Parfum intense » fait référence au concept d'intensité du parfum ;
- « un très bon sillage » fait référence au concept de sillage du parfum ;
- « Il tient toute la journée. » fait référence au concept de tenue ;
- « Il est parfait aussi pour une soirée » fait référence à une activité sociale ;
- « Produit efficace qui respecte la nature », « Issu de techniques respectueuses de l'environnement et produit par une société locale », et « bougie d'ambiance assez insatisfaisante compte tenu de sa composition generant une pollution de l'intérieur du ligement » font référence aux caractéristiques du produit ;
- « Je ne peux que recommander ce produit » fait référence à la recommandation du produit ;
- « Le prix est un peu cher » fait référence au concept de prix.

Le fait qu'un commentaire aborde certaines thématiques plutôt que d'autres pourrait avoir un lien avec les valeurs de l'auteur du commentaire. C'est notre hypothèse actuelle et nous allons la tester au cours de notre recherche.

Les thématiques que nous avons extraites ci-dessus sont basées sur une ontologie que nous avons créée et que nous présentons rapidement dans le paragraphe suivant.

4.2. Ontologie

L'idée de cette ontologie est centrée sur les individus et les produits parfumés : un individu vit dans un environnement et interagit avec la société qui l'entoure. Un individu achète et utilise des produits. Un produit peut donner des effets fonctionnels et émotionnels. Il a aussi des attributs : emballage, taille, prix, et fragrance. La fragrance peut être décrite sous plusieurs dimensions : sa tenue, son intensité, sa diffusivité, etc. Nous nous sommes inspirés des études sur les parfums du domaine de la psychologie pour construire la partie sur la fragrance de cette ontologie (Manetta et al., 2007).

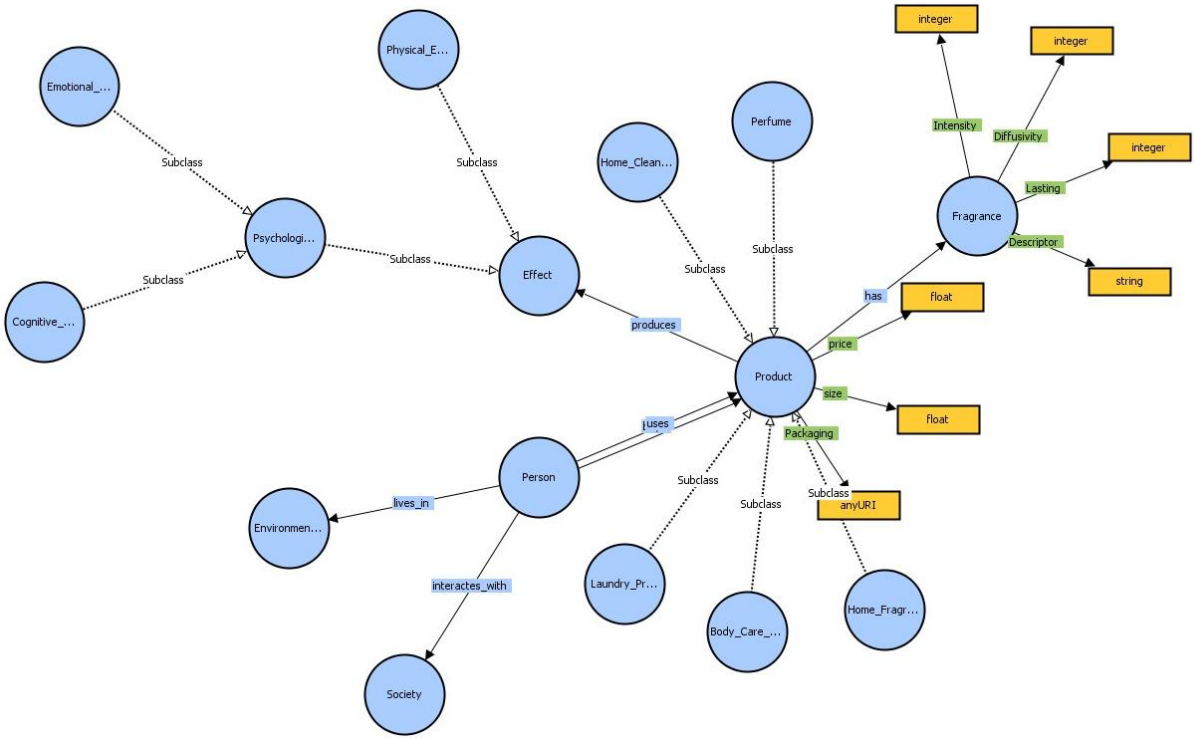

FIGURE 3. Ontologie sous la forme de graphe

5. Conclusion

Au cours de notre projet de recherche, nous allons nous inspirer des différentes méthodes de la détection de personnalité citées dans la partie 3, et tester certaines d'entre elles pour notre objectif de détection des signaux des valeurs humaines. En plus de ces méthodes, ce projet de recherche nous donne aussi l'occasion de tester la performance des modèles pré-entraînés de grande taille comme le GPT-3 (Brown et al., 2020) pour des tâches abstraites (en l'occurrence psycholinguistique).
Pour atteindre cet objectif, nous allons aussi construire un outil de l'analyse textuelle adapté au domaine de la parfumerie. Cela nous donnera l'occasion d'examiner si les connaissances sur la parfumerie peuvent contribuer à la performance de détection des valeurs dans le texte.

Remerciement

Ce travail est effectué dans le cadre d'une convention CIFRE, gérée par l'Association Nationale de la Recherche Technique (ANRT), et établie entre le Laboratoire ERTIM de l'Inalco et la société IFF. Un grand merci à Dr Frédérique SEGOND et à Dr Céline MANETTA pour leur relecture de l'article et leur soutien au projet de recherche.

Références

ARIJANTO J. E., GERALDY S., TANIA C., & SUHARTONO D. (2021). Personality Prediction Based on Text Analytics Using Bidirectional Encoder Representations from Transformers from English Twitter Dataset. International Journal of Fuzzy Logic and Intelligent Systems, 21(3), 310-316.

BLEI D. M., NG A. Y., & JORDAN M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., ... & AMODEI D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

CAMBRIA E., HAVASI C., & HUSSAIN A. (2012). Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In Twenty-Fifth international FLAIRS conference.

CHENG A. S., & FLEISCHMANN K. R. (2010). Developing a meta-inventory of human values. Proceedings of the American Society for Information Science and Technology, 47(1), 1-10.

COLTHEART M. (1981). The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology Section A, 33(4), 497-505.

DARLIANSYAH A., NAEEM M. A., MIRZA F., & PEARS R. (2019). SENTIPEDE: A Smart System for Sentiment-based Personality Detection from Short Texts. J. Univers. Comput. Sci., 25(10), 1323-1352.

DEVLIN J., CHANG M. W., LEE K., & TOUTANOVA K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

DIGMAN, J. M. (1990). Personality structure: Emergence of the five-factor model. Annual review of psychology, 41(1), 417-440.
Havasi C., Speer R., & Alonso J. (2007). ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In Recent advances in natural language processing (pp. 27-29).
HAVASI C., SPEER R., & PUSTEJOVSKY J. (2009). Automatically suggesting semantic structure for a generative Lexicon ontology. Generative Lexicon.
HERNANDEZ R. K., & SCOTT I. (2017). Predicting Myers-Briggs type indicator with text. In 31st Conference on Neural Information Processing Systems (NIPS 2017).
ISHITA E., OARD D. W., FLEISCHMANN K. R., CHENG A. S., & TEMPLETON T. C. (2010). Investigating multi-label classification for human values. Proceedings of the American Society for Information Science and Technology, 47(1), 1-4.
IVANOV A. V., RICCARDI G., SPORKA A. J., & FRANC J. (2011). Recognition of personality traits from human spoken conversations. In *Twelfth Annual Conference of the International Speech Communication Association*.
JOHN O. P., & SRIVASTAVA S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality: Theory and research, 2(1999), 102-138.
KEH S. S., & CHENG I. (2019). Myers-Briggs personality classification and personality-specific language generation using pre-trained language models. arXiv preprint arXiv:1907.06333.
LEONARDI S., MONTI D., RIZZO G., & MORISIO M. (2020). Multilingual transformer-based personality traits estimation. Information, 11(4), 179.
LI M. LIN Y. HOOVER J. WHITEHEAD S. VOSS C. DEHGHANI M. AND JI H. (2019). Multilingual entity, relation, event and human value extraction. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) (pp. 110-115).
LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., ... & STOYANOV V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
LUK C. L., & BOND M. H. (1993). Personality variation and values endorsement in Chinese university students. *Personality and Individual Differences*, *14*(3), 429-437.
MAIRESSE F., WALKER M. A., MEHL M. R., & MOORE R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of artificial intelligence research, 30, 457-500.
MAJUMDER N., PORIA S., GELBUKH A., & CAMBRIA E. (2017). Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems, 32(2), 74-79.
MANETTA C., SANTARPIA A., SANDER E., MONTET A., & URDAPILLETA I. (2007). Catégorisation du langage descriptif et du langage figuré dans l'expérience des parfums complexes. Psychologie française, 52(4), 479-497.
MIKOLOV T., CHEN K., CORRADO G., & DEAN J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
MYERS I. B., MCCAULLEY M. H., QUENK N. L., & HAMMER A. L. (1998). MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator. Consulting Psychologists Press.
PARK G., SCHWARTZ H. A., EICHSTAEDT J. C., KERN M. L., KOSINSKI M., STILLWELL D. J., ... & SELIGMAN M. E. (2015). Automatic personality assessment through social media language. Journal of personality and social psychology, 108(6), 934.
PENNEBAKER J. W., & KING L. A. (1999). Linguistic styles: language use as an individual difference. Journal of personality and social psychology, 77(6), 1296.
PENNEBAKER J. W., FRANCIS M. E., & BOOTH R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.
PENNEBAKER J. W., BOYD R. L., JORDAN K., & BLACKBURN K. (2015). The development and psychometric properties of LIWC2015.
PENNINGTON J., SOCHER R., & MANNING C. D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
PLANK B., & HOVY D. (2015, September). Personality traits on twitter—or—how to get 1,500 personality tests in a week. In Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 92-98).
PORIA S., GELBUKH A., HUSSAIN A., HOWARD N., DAS D., & BANDYOPADHYAY S. (2013). Enhanced SenticNet with affective labels for concept-based opinion mining. IEEE Intelligent Systems, 28(2), 31-38.
PORIA S., GELBUKH A., AGARWAL B., CAMBRIA E., & HOWARD N. (2013). Common sense knowledge based personality recognition from text. In Mexican International Conference on Artificial Intelligence (pp. 484-496). Springer, Berlin, Heidelberg.
REN Z., SHEN Q., DIAO X., & XU H. (2021). A sentiment-aware deep learning approach for personality detection from text. Information Processing & Management, 58(3), 102532.
ROCCAS S., SAGIV L., SCHWARTZ S. H., & KNAFO A. (2002). The big five personality factors and personal values. *Personality and social psychology bulletin*, *28*(6), 789-801.
SCHWARTZ S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. Advances in experimental social psychology, 25(1), 1-65.
SCHWARTZ S. H. (1996). Value Priorities and Behavior: Applying a theory of integrated value systems. In C. Seligman, J.M. olson & M.P Zanna (Eds.), The psychology of values: The Ontario symposium, volume 8(pp. 1-24). Manwah, N.J.: Erlbaum.
SCHWARTZ S. H., MELECH G., LEHMANN A., BURGESS S., HARRIS M., & OWENS V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. Journal of cross-cultural psychology, 32(5), 519-542.
SCHWARTZ S. H. (2003). A proposal for measuring value orientations across nations. Questionnaire Package of the European Social Survey. 259-290.
SCHWARTZ S. H. (2006). Les valeurs de base de la personne : théorie, mesures et applications. Revue française de sociologie. 47 (4). 929-968.
ŠTAJNER S., & YENIKENT S. (2021). Why Is MBTI Personality Detection from Texts a Difficult Task?. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 3580-3589).
STILLWELL D. J., AND KOSINSKI M. (2011) mypersonality research wiki
TAKAYAMA Y., TOMIURA Y., ISHITA E., OARD D. W., FLEISCHMANN K. R., & CHENG A. S. (2014). A word-scale probabilistic latent variable model for detecting human values. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (pp. 1489-1498).
TAUSCZIK Y. R., & PENNEBAKER J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.
VÁSQUEZ R. L., & OCHOA-LUNA J. (2021, October). Transformer-based Approaches for Personality Detection using the MBTI Model. In 2021 XLVII Latin American Computing Conference (CLEI) (pp. 1-7). IEEE.
VERPLANKEN B., & HOLLAND R. W. (2002). Motivated decision making: effects of activation and self-centrality of values on choices and behavior. Journal of personality and social psychology, 82(3), 434.
YANG Z., DAI Z., YANG Y., CARBONELL J., SALAKHUTDINOV R. R., & LE Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.