Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Les avis sur les restaurants à l'épreuve de l'apprentissage automatique

Hyun Jung Kang Iris Eshkol-Taravella
MoDyCo UMR7114, 200 Avenue de la République, 92001 Nanterre, France
hyunjung.kang@parisnanterre.fr, ieshkolt@parisnanterre.fr

RÉSUMÉ

Dans la fouille d'opinions, de nombreuses études portent sur l'extraction automatique des opinions positives ou négatives. Cependant les recherches ayant pour objet la fouille de suggestions et d'intentions sont moins importantes, malgré leur lien profond avec l'opinion. Cet article vise à détecter six catégories (opinion positive/mixte/négative, suggestion, intention, description) dans les avis en ligne sur les restaurants en exploitant deux méthodes : l'apprentissage de surface et l'apprentissage profond supervisés. Les performances obtenues pour chaque catégorie sont interprétées ensuite en tenant compte des spécificités du corpus traité.

ABSTRACT

An Empirical Examination of Online Restaurant Reviews

In opinion mining, many works focus on the automatic extraction of positive or negative opinions. However, researches on suggestion and intention mining haven't been addressed much, despite their strong connection to opinion. This article aims to detect six categories (positive/negative/mixed opinion, suggestion, intention, description) in online restaurant reviews using two methods : traditional supervised learning and deep learning. We then interpret the performances obtained for each category by taking into account the specificities of the corpus treated.

MOTS-CLÉS : fouille d'opinions, avis en ligne, apprentissage supervisé, apprentissage profond, suggestion, intention.

KEYWORDS: opinion mining, online reviews, machine learning, deep learning, suggestion, intention.

1 Introduction

Les avis en ligne sont des sources d'analyses dans différents domaines : le marketing, l'informatique, la linguistique, le TAL où ils sont souvent considérés comme relevant de la notion d'opinion. De nombreux travaux portent sur la fouille d'opinions dont l'objectif est de classifier un document (Pang *et al.*, 2002; Turney, 2002) ou une phrase (Wiebe *et al.*, 1999) selon sa polarité. L'opinion peut être aussi extraite comme un tuple (ABSA ; *Aspect Based Sentiment Analysis*) qui se compose des éléments suivants : entité, aspect, sentiment, porteur et temps (Hu & Liu, 2004; Liu, 2012; Hamon *et al.*, 2015; Lark, 2017).

D'autres notions apparaissent en lien avec les avis comme les suggestions ou les intentions. (Benamara *et al.*, 2017), par exemple, affirment que la détection des intentions permet de compléter l'analyse de

Volume 2 : Traitement Automatique des Langues Naturelles, pages 249–257.
HAL : hal-02784774.

sentiments et d'opinions. Cependant les recherches ayant pour objet la détection de suggestions et d'intentions sont moins nombreuses. Le travail de (Ramanand *et al.*, 2010) est, selon nous, le premier à s'intéresser à la détection de suggestions. Il distingue deux types de souhaits (*wish*) : le souhait d'améliorer un produit et le souhait d'acheter celui-ci. (Brun & Hagège, 2013) analysent les avis sur les produits et établissent un ensemble de règles de leur détection en se fondant sur des éléments linguistiques. (Negi & Buitelaar, 2015) étudient les avis sur les hôtels et les produits électroniques dans lesquels les auteurs formulent des conseils ou offrent des suggestions aux futurs utilisateurs. (Negi *et al.*, 2016) évaluent différentes méthodes de détection de suggestion telles que les règles linguistiques élaborées manuellement, les machines vectorielles de support (SVM) et l'apprentissage profond. L'une des tâches proposées à l'atelier SemEval-2019 (Negi *et al.*, 2019) avait pour objectif d'extraire les suggestions dans les avis et les forums sur Internet. La détection d'intentions est abordée dans le travail de (Carlos & Yalamanchi, 2012) qui catégorise l'intention dans le domaine du marketing et du service clientèle. (Chen *et al.*, 2013) effectuent une classification des intentions explicites. (Ding *et al.*, 2015) proposent un modèle, basé sur les réseaux de neurones convolutifs (CNN, *Convolutional Neural Network*), pour identifier si l'utilisateur manifeste une intention de consommation.

Les avis traités dans ce travail concernent les restaurants. Leur analyse linguistique a permis de proposer un modèle conceptuel qui dépasse la notion d'opinion positive/négative/mixte et qui intègre une dimension linguistique (Eshkol-Taravella & Kang, 2019). Le modèle propose trois nouvelles classes (i.e. suggestion, intention, description) qui sont moins étudiées dans la fouille d'opinion.

La suggestion est un conseil émis par un visiteur. Ses marqueurs linguistiques sont les verbes de parole comme « recommander », « conseiller », le mode conditionnel et impératif, les pronoms personnels (« vous ») et les adjectifs possessifs de la deuxième personne (« votre », « vos »). Les suggestions peuvent être adressées à la fois aux restaurants (afin qu'ils prennent conscience des problèmes) et aux autres clients potentiels (futurs visiteurs) comme dans les exemples suivants : « Je conseille le tiramisu. », « Une lumière un peu plus tamisée aurait été parfaite ». Pourtant, les travaux précédents consacrés à leur détection ne prennent en compte que l'un des deux destinataires.

L'intention est un souhait exprimé explicitement de revenir ou de ne pas pas revenir dans un restaurant, elle montre un engagement volontaire du visiteur (« On reviendra ! », « Nous ne reviendrons pas ! »). L'intention est marquée de manière explicite à travers les pronoms « je », « nous » et « on », les verbes au futur et le préfixe verbal d'itération « re- ».

La description concerne les informations factuelles associées à l'expérience vécue comme « Soirée pour notre anniversaire de mariage » ou « Nous y étions un midi », qui sont peu reconnues dans la fouille d'opinion. Malgré sa nature objective et sa faible fréquence, la description est néanmoins une information intéressante car elle permet aux lecteurs de découvrir l'arrière-plan de l'expérience comme la raison pour laquelle les visiteurs se rendent dans le restaurant, les personnes qui les accompagnent, *etc*.

(Eshkol-Taravella & Kang, 2019) ont présenté la détection automatique de ces catégories fondée sur l'apprentissage supervisé en comparant différents modèles (*Naïve Bayes, Support Vector Machine, Logistic Regression*) tout en tenant compte du déséquilibre entre les classes d'évaluation créées. Le score F-mesure 0,88 a été obtenu en utilisant le sur-échantillonnage de ADASYN associé à l'algorithme SVM. Le travail présenté ici vise deux objectifs : (1) comparer les résultats de deux techniques d'apprentissage supervisé utilisées (l'apprentissage de surface et l'apprentissage profond) ; (2) examiner et comparer les performances obtenues pour chaque catégorie détectée. La méthodologie appliquée est décrite dans la section 2. Les expériences de détection automatique des six catégories

proposées exploitant les techniques de l'apprentissage de surface et de l'apprentissage profond ainsi que l'analyse de leurs performances sous diverses facettes sont présentées dans la section 3.

2 Méthodologie

2.1 Annotations et prétraitements

Données traitées. Nous avons collecté 21 158 avis sur 87 restaurants situés à Paris depuis un site internet [1] dont 6 287 avis (17 268 phrases, avec une moyenne de dix mots par phrase) ont été annotées. Les prétraitements réalisés sont décrits dans (Eshkol-Taravella & Kang, 2019). Nous nous contentons de présenter ici leur synthèse.

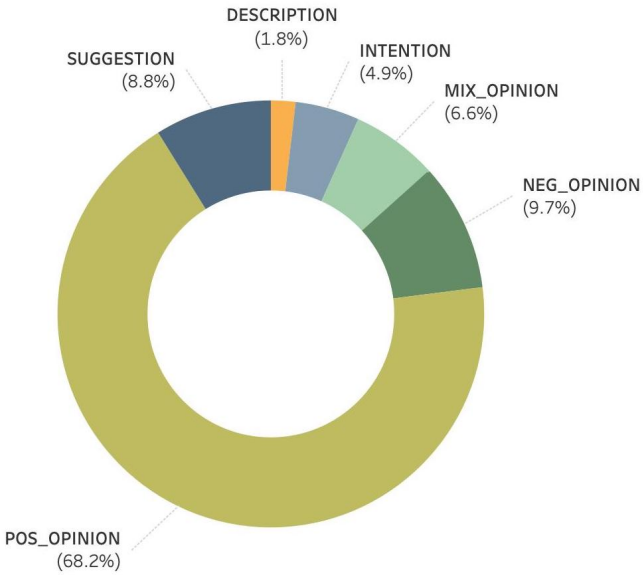

FIGURE 1 – Répartition des classes

Annotation. Chaque phrase a été annotée selon l'une des six catégories : POS_OPINION, NEG_OPINION, MIX_OPINION, SUGGESTION, INTENTION et DESCRIPTION. Dans les cas d'ambiguïté où plusieurs catégories sont possibles, SUGGESTION ou INTENTION ont été privilégiées car celles-ci sont peu représentées dans les données. La phrase « Toujours aussi excellent, nous y retournerons c'est certain » par exemple, peut être classée dans les deux catégories POS_OPINION et INTENTION. Nous l'avons cependant annotée comme INTENTION. La tâche d'annotation a été évaluée par trois annotateurs (i.e. les doctorants en linguistique). Selon la mesure Kappa de Fleiss, nous avons obtenu 0,90, un accord considéré comme « presque parfait » (Landis & Koch, 1977). La distribution des catégories dans le corpus annoté s'est avérée non homogène (voir figure 1) : l'étiquette POS_OPINION constitue la majorité des catégories, représentant 68,2%, alors que DESCRIPTION et INTENTION n'en font que 1,8% et 4,9%.

Normalisation. Pour normaliser et nettoyer le corpus, certains traitements ont été effectués : le remplacement des émoticônes par « emoPOS » ou « emoNEG » selon la polarité [2], le passage des mots en

1. La Fourchette, https://www.lafourchette.com/
2. Les émoticônes se distinguent généralement par la polarité : ceux qui ont une polarité positive comme « :) », « :-) »

minuscules ; l'élimination de la ponctuation [3] ; la normalisation des mots consistant en la transformation des abréviations par leurs variations complètes (e.g. « resto » par « restaurant ») ; le remplacement des chiffres par une étiquette « NUM » [4] ; la lemmatisation en utilisant StanfordCoreNLP [5].

2.2 Classification

La classification automatique des phrases dans les six catégories prédéfinies a été effectuée en utilisant trois méthodes : SVM (*Support Vector Machine*) linéaire, CNN (*Convolutional Neural Network*) et LSTM (*Long Short-Term Memory network*). Pour les expériences basées sur SVM, nous avons utilisé la librairie scikit-learn [6] et pour celles de CNN et LSTM, nous avons employé la librairie Keras [7] avec TensorFlow [8]. Pour évaluer toutes les expériences, une validation croisée stratifiée à 5 plis a été effectuée. Les techniques pour gérer l'équilibre de la répartition des catégories n'ont pas été appliquées dans cette étude.

SVM linéaire. Deux approches de représentation de texte ont été exploitées : l'approche du sac de mots et celle du plongement de mots (*embedding*). Pour la première approche, nous avons supprimé les mots-outils [9] et employé *CountVectorizer* et *TfidfVectorizer* de scikit-learn. Nous avons testé deux paramètres dont ces méthodes disposent, *n_gram* et *max_feature*. La meilleure combinaison des paramètres a été obtenue avec une procédure de grille de recherche (*GridSearch* [10]).

La deuxième approche concerne le plongement de mots, méthode améliorée par rapport à celle du sac de mots car capable de prendre en compte les similarités contextuelles entre les mots. Pour ce faire, nous avons entraîné Word2vec (Mikolov *et al.*, 2013) du type de CBOW (sac de mots continus) avec Gensim [11]. La taille de la fenêtre a été fixée à 6. Les traits (*features*) pris en compte lors de l'apprentissage sont : les catégories morphosyntaxiques jugées pertinentes (e.g. les noms, les verbes, les adjectifs) proposées par StanfordCoreNLP, les différentes variations des verbes (e.g. les verbes au futur, les verbes au conditionnel, les verbes possédant le préfixe 're-'), la négation, les mots positifs et négatifs [12], les scores de polarité et de subjectivité [13], le connecteur « mais », le symbole € (*euro*), les chiffres, les émoticônes, les multiples ponctuations en cascade, les mots en majuscule [14], la longueur de la phrase, la diversité et la densité lexicales. En nous basant sur ces traits définis, l'algorithme SVM linéaire a été appliqué.

CNN. Pour exploiter la technique des CNN, nous nous sommes appuyés sur la proposition de configurations proposée dans (Zhang & Wallace, 2017) pour une tâche de classification automatique

et ceux qui sont plutôt de polarité négative « :(». Une liste comprenant les émoticônes positifs et négatifs a été créée préalablement.

3. Pour éviter des interférences entre les traitements réalisés, leur ordre a été prédéfini.
4. Les informations factuelles (le nombre de personnes, l'heure, le prix, etc.) sont souvent présentées en chiffres.
5. https://stanfordnlp.github.io/CoreNLP/download.html
6. http://scikit-learn.org/stable/
7. https://keras.io/
8. https://www.tensorflow.org/
9. La liste des mots-outils proposée par NLTK (https://www.nltk.org/) a été modifié.
10. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
11. https://radimrehurek.com/gensim/ ; Les modèles Word2vec entraînés sur le corpus frWiki (Wikipédia) par (Fauconnier, 2015) ont été testés, or, le résultat des modèles pré-entraînés a été moins intéressant que nos propres modèles.
12. Proposés par Textblob (https://textblob.readthedocs.io/en/dev/)
13. Obtenus par Textblob
14. Puisque les données ont été normalisées, la fréquence des multiples ponctuations en cascade et des mots en majuscule sont extraites des données originales.

des textes. Nous avons pris en entrée une matrice d'embedding à l'aide de Word2vec entraîné précédemment sur notre corpus. Par la suite, nous avons appliqué un filtre de convolution de 32 neurones ainsi qu'un kernel de taille 4 suivie par la fonction d'activation ReLU (Unité de Rectification Linéaire). Une couche de *Max pooling* a ensuite été appliquée sur la sortie de la couche de convolution, divisant par 2 la sortie de la couche précédente. Enfin, nous avons exploité l'aplatissement et la réduction de dimension, en appliquant une couche dense à 10 unités, suivie par la fonction d'activation ReLU, avant d'exploiter une activiation softmax sur la couche finale composée de 6 neurones.

LSTM. Les LSTMs sont une variante de réseaux de neurones récurrents (RNN) considérés comme très performantes sur de longues séquences (Osinga, 2018). Nous avons utilisé comme entrée la même matrice d'embeddings que celui de CNN, suivie d'une couche avec 100 unités, envoyée ensuite à une couche dense et finissante par une activation softmax. L'avantage des LSTMs est de mieux prendre en compte les dépendances entres mots distants.

Pour ces types de réseaux de neurons, CNN et LSTM, les hyperparamètres choisis sont l'optimiseur Adam (*Adaptive Moments*) et une perte d'entropie. La taille du batch est de 5, avec 7 époques.

3 Résultats

Les performances des différents modèles ont été évaluées en calculant pour chacun d'eux la moyenne pondérée de la précision, du rappel, de la F-mesure et la matrice de confusion. La macro F-mesure, donnant un poids identique à chaque catégorie, ne tient pas compte de la répartition déséquilibrée des classes. Étant donné que cette répartition est asymétrique dans les données, la moyenne pondérée de la F-mesure est jugée pertinente [15], c'est la raison pour laquelle cette mesure est utilisée pour l'évaluation. Le tableau 1 illustre la comparaison des performances entre les méthodes employées. Le SVM linéaire utilisé avec le sac de mots (i.e. l'apprentissage de surface) a donné le meilleur résultat avec une moyenne pondérée de F-mesure égale à 0,88. L'apprentissage de surface a donc produit dans le cadre de cette étude un meilleur résultat que l'apprentissage profond. Nous considérons pourtant que ce dernier peut encore être amélioré au moyen de réseaux de neurones plus complexes. Bien que Word2vec soit présenté comme une approche optimale, nos résultats montrent que l'approche du sac de mots peut s'avérer plus performante que Word2vec dans les cas où les données ne sont pas importantes.

sac de mots+linearSVM	Word2vec+linearSVM	CNN	LSTM
0.88	0.80	0.85	0.84

TABLE 1 – La comparaison des moyennes pondérées de la F-mesure entre les méthodes employées

La matrice de confusion offre une vision globale de la meilleure performance, comme le montre la figure 2. Chaque ligne correspond à la classe réelle et chaque colonne à la classe prédite. Les cellules de la diagonale principale indiquent celles qui sont classifiées correctement. La cellule DESCRIPTION est légèrement plus claire car l'information réunie sous cette étiquette était plus difficile à être classifiée. Ce constat peut s'expliquer en partie par le manque d'échantillons de DESCRIPTION et donc par le faible nombre de traits fournis durant l'apprentissage. Par ailleurs, cette catégorie est très hétérogène et varie en fonction du profil de l'internaute, ce qui donne lieu à un

15. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score

large éventail de vocabulaires et de contextes. Par conséquent, DESCRIPTION a une tendance à être rangée dans la classe majoritaire, c'est-à-dire POS_OPINION (0,41) et occasionnellement comme NEG_OPINION (0,14). Nous observons également une tendance similaire pour MIX_OPINION, dont le mauvais score (0,66) vient du fait que la catégorie implique à la fois POS_OPINION (0,19) et NEG_OPINION (0,12) [16].

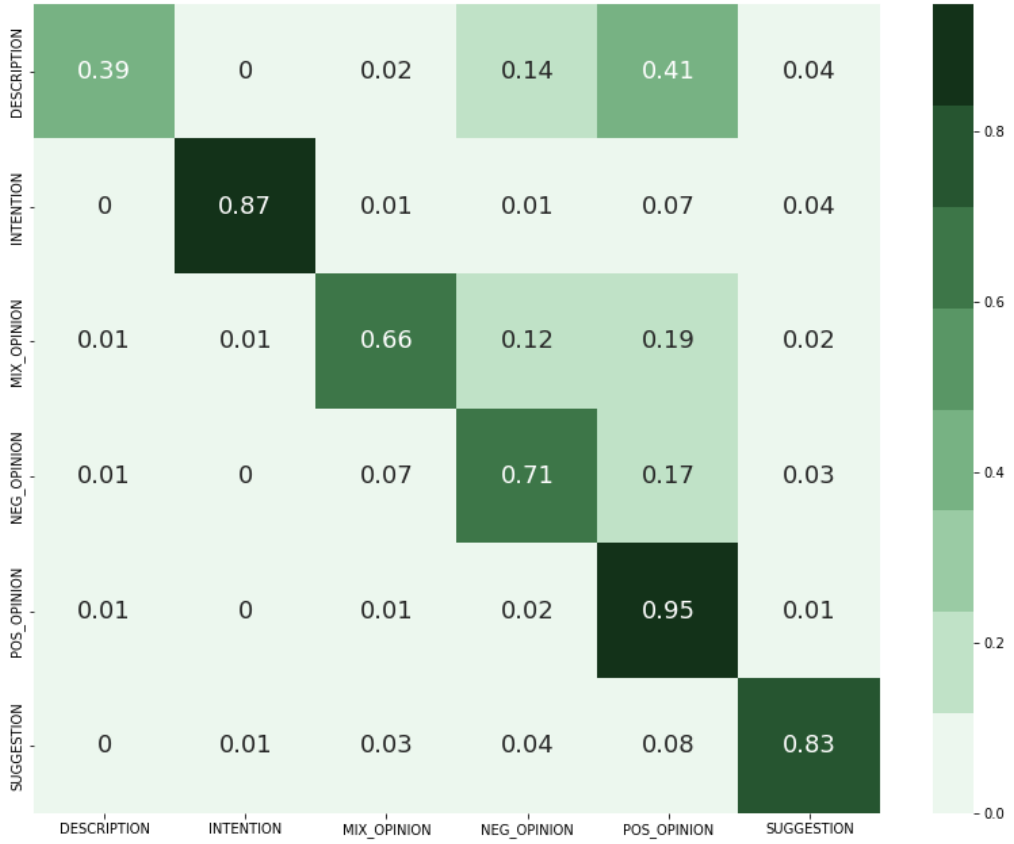

FIGURE 2 – La matrice de confusion normalisée de SVM linéaire utilisé avec le sac de mots

La figure 3 présente une comparaison de la moyenne pondérée de la précision, du rappel et de la F-mesure entre l'apprentissage de surface (sac de mots+SVM linéaire) et l'apprentissage profond (CNN) pour chaque catégorie d'évaluation. Une étiquette POS_OPINION est détectée avec la meilleure performance (la F-mesure étant d'environ 0,94) et l'INTENTION, obtient le deuxième meilleur score (0,86-0,88). Une étiquette DESCRIPTION est détectée avec la plus mauvaise performance (0,34-0,46) ayant le plus grand écart entre la précision et le rappel ce qui correspond à 0,16 pour chaque méthode d'apprentissage. Ce résultat est dû à sa faible fréquence dans le corpus. Par ailleurs, elle semble s'appuyer sur peu de marqueurs lexicaux car sa nature est très hétérogène.

D'une manière générale, les performances de l'apprentissage de surface sont supérieures à celle de l'apprentissage profond dans la majorité des cas. Les étiquettes DESCRIPTION et SUGGESTION ont une tendance à être mieux détectées avec l'apprentissage de surface ce qui montre que les traits linguistiques retenus sont utiles dans l'apprentissage. L'INTENTION a une précision supérieure mais un rappel inférieur à la précision lorsque l'apprentissage profond est exploité, donnant au final une F-mesure similaire à l'apprentissage de surface.

16. Par exemple, « Vraiment très bien, juste un peu trop bruyant. ».

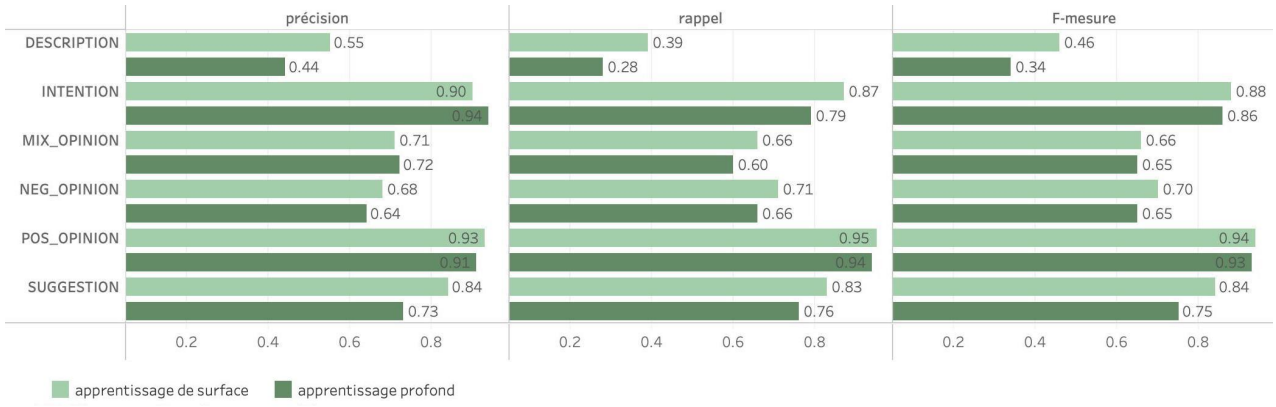

FIGURE 3 – Précision, Rappel et F-mesure d'apprentissage supervisé (sac de mots+linearSVC) et d'apprentissage profond (CNN)

Une catégorie MIX_OPINION semble poser également des difficultés aux classifieurs. L'observation manuelle du corpus montre que le changement de polarité est marqué souvent par les conjonctions comme « mais » (e.g. « Le restaurant était complet et très bruyant mais la cuisine excellente. », « Beau restaurant, très bons plats mais service à revoir. »). La segmentation du corpus selon les conjonctions pourrait améliorer la performance de la détection de cette catégorie mais aussi pourrait permettre sa suppression. La segmentation de la phrase « Le restaurant était complet et très bruyant mais la cuisine excellente. » en deux propositions « Le restaurant était complet et très bruyant » et « la cuisine excellente » permet de classer la première proposition dans une catégorie NEG_OPINION et la seconde dans POS_OPINION. L'approche de (Lark, 2017), qui emploie les règles lexicosyntaxiques appliquées sur les connecteurs, semble pouvoir résoudre cette difficulté.

4 Conclusion

Cet article décrit l'expérience portant sur la classification automatique des avis selon six catégories prédéfinies en exploitant deux techniques de l'apprentissage supervisée : l'apprentissage de surface et l'apprentissage profond. L'approche de surface obtient la meilleure moyenne pondérée de la F-mesure (0,88), qui chute légèrement dans le cas de l'apprentissage profond (0,86). Notons que l'apprentissage de surface est plus coûteux en termes de temps et d'efforts à cause des traits à fournir aux algorithmes. Lorsque l'approche de surface est appliquée, les moyennes pondérées de la F-mesure pour chaque catégorie sont : POS_OPINION (0,94), INTENTION (0,88), SUGGESTION (0,84), NEG_OPINION (0,70), MIX_OPINION (0,66) et DESCRIPTION (0,46). Parmi les trois nouvelles catégories proposées : INTENTION, SUGGESTION, DESCRIPTION, c'est la détection de cette dernière qui obtient les résultats moins satisfaisants. Pour améliorer son score, il faudrait augmenter la taille du corpus de référence. Par ailleurs, il serait intéressant de mesurer la généricité des catégories INTENTION et SUGGESTION dans d'autres corpus.

Références

BENAMARA F., TABOADA M. & MATHIEU Y. (2017). Evaluative language beyond bags of words :

Linguistic insights and computational applications. *Computational Linguistics*, **43**(1), 201–264. DOI : 10.1162/COLI_a_00278.

BRUN C. & HAGÈGE C. (2013). Suggestion mining : Detecting suggestions for improvement in users' comments. *Research in Computing Science*, **70**, 199–209.

CARLOS C. S. & YALAMANCHI M. (2012). Intention analysis for sales, marketing and customer service. In *Proceedings of COLING 2012 : Demonstration Papers*, p. 33–40, Mumbai, India : The COLING 2012 Organizing Committee.

CHEN Z., LIU B., HSU M., CASTELLANOS M. & GHOSH R. (2013). Identifying intention posts in discussion forums. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1041–1050, Atlanta, Georgia : Association for Computational Linguistics.

DING X., LIU T., DUAN J. & NIE J.-Y. (2015). Mining user consumption intention from social media using domain adaptive convolutional neural network. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

ESHKOL-TARAVELLA I. & KANG H. J. (2019). Observation de l'expérience client dans les restaurants. In *TALN 2019*.

FAUCONNIER J.-P. (2015). French word embeddings. http://fauconnier.github.io.

HAMON T., FRAISSE A., PAROUBEK P., ZWEIGENBAUM P. & GROUIN C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (deft).

HU M. & LIU B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 168–177.

LANDIS J. R. & KOCH G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.

LARK J. (2017). *Construction semi-automatique de ressources pour la fouille d'opinion*. Thèse de doctorat, Nantes.

LIU B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv : 1301.3781.

NEGI S., ASOOJA K., MEHROTRA S. & BUITELAAR P. (2016). A study of suggestions in opinionated texts and their automatic detection. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, p. 170–178, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/S16-2022.

NEGI S. & BUITELAAR P. (2015). Towards the extraction of customer-to-customer suggestions from reviews. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. DOI : 10.18653/v1/D15-1258.

NEGI S., DAUDERT T. & BUITELAAR P. (2019). Semeval-2019 task 9 : Suggestion mining from online reviews and forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, p. 783–883.

OSINGA D. (2018). *Deep Learning Cookbook*. Practical Recipes to Get Started Quickly. O'Reilly Media.

PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? sentiment classification using machine learning techniques. *EMNLP*, **10**. DOI : 10.3115/1118693.1118704.

RAMANAND J., BHAVSAR K. & PEDANEKAR N. (2010). Wishful thinking : Finding suggestions and 'buy' wishes from product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, p. 54–61, Stroudsburg, PA, USA : Association for Computational Linguistics.

TURNEY P. D. (2002). Thumbs up or thumbs down ? : Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 417–424, Stroudsburg, PA, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073153.

WIEBE J. M., BRUCE R. F. & O'HARA T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, p. 246–253, Stroudsburg, PA, USA : Association for Computational Linguistics. DOI : 10.3115/1034678.1034721.

ZHANG Y. & WALLACE B. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 253–263, Taipei, Taiwan : Asian Federation of Natural Language Processing.