

# Caractéristiques linguistiques pour la prédiction de la difficulté des phrases dans l'analyse de sentiment basée sur les aspects

Adrian-Gabriel Chifu<sup>1,\*</sup>, Sébastien Fournier<sup>1,†</sup>

<sup>1</sup>Aix-Marseille Université, Université de Toulon, CNRS, LIS, France

## Abstract

Ceci est un résumé en français de l'article intitulé "Linguistic features for sentence difficulty prediction in ABSA" [1], accepté pour publication à la conférence ICIRA, consultable sur : <https://arxiv.org/abs/2402.03163>.

## Keywords

Analyse de sentiments, Difficulté, Classification, Apprentissage automatique

## 1. Introduction

L'analyse des sentiments est un domaine qui vise à extraire et analyser ces éléments subjectifs du texte, et elle peut être appliquée à différents niveaux de granularité, tels que le document, le paragraphe, la phrase ou l'aspect. L'analyse de sentiment basée sur les aspects (ABSA) constitue un champ d'étude majeur dans le traitement automatique du langage naturel, visant à identifier et analyser les opinions exprimées sur des aspects spécifiques d'un produit ou service dans des textes [2]. Malgré les avancées significatives, la tâche d'ABSA reste complexe en raison de la diversité linguistique, de l'ambiguïté et de la subjectivité des expressions sentimentales. Cette recherche vise à explorer un aspect relativement peu étudié dans l'ABSA : la prédiction de la difficulté des phrases. Trois questions de recherche guident notre étude :

- **QR1** : Quel impact la diversité thématique et syntaxique a-t-elle sur la difficulté de l'analyse de sentiment basée sur les aspects ?
- **QR2** : Quelles caractéristiques linguistiques sont les plus pertinentes pour estimer la difficulté des phrases dans l'ABSA ?
- **QR3** : Quelle est la différence de performance entre les modèles traditionnels de machine learning et les grands modèles de langage dans la prédiction de la difficulté des phrases ?

Pour répondre à ces questions, nous avons mené des expériences sur trois jeux de données annotés pour l'ABSA ("Laptops", "Restaurants" [3], et "MTSC" [4]), et une version fusionnée de

---

CORIA 2024 (Conférence en Recherche d'Information et Applications), 3-4 avril 2022, La Rochelle, France


\*Auteur correspondant.

† Ces auteurs ont contribué de manière égale.

✉ [adrian.chifu@univ-amu.fr](mailto:adrian.chifu@univ-amu.fr) (A. Chifu); [sebastien.fournier@univ-amu.fr](mailto:sebastien.fournier@univ-amu.fr) (S. Fournier)

🌐 <https://adrianchifu.com> (A. Chifu)

🆔 0000-0003-4680-5528 (A. Chifu)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

ces trois, afin d'examiner l'impact de la diversité sur la difficulté de la tâche.

## 2. Méthodologie

Nous étudions l'impact de la diversité des domaines et de la diversité syntaxique sur la difficulté. Nous utilisons une combinaison de classificateurs pour identifier les phrases les plus difficiles et analyser leurs caractéristiques. Nous avons adopté deux définitions pour mesurer la difficulté des phrases : une approche binaire et une approche à six niveaux, basées sur la capacité des classificateurs à prédire correctement la polarité du sentiment. Neuf caractéristiques linguistiques ont été sélectionnées pour prédire la difficulté, incluant des éléments syntaxiques et sémantiques.

Nous avons analysé la performance de 21 modèles de classification, incluant des approches traditionnelles et des Transformeurs comme BERT [5], pour évaluer la difficulté des phrases dans nos ensembles de données.

## 3. Résultats

Les résultats montrent que la prédiction de la difficulté dans l'ABSA est possible mais dépend fortement des ensembles de données et de la définition de la difficulté. Les caractéristiques linguistiques telles que la présence d'entités nommées et la complexité syntaxique semblent influencer significativement la difficulté.

Les Transformeurs, en particulier BERT, ont montré une meilleure performance dans les ensembles de données diversifiés, soulignant leur potentiel pour gérer la complexité inhérente à l'ABSA.

## 4. Conclusion

Cette étude apporte des éclaircissements sur les défis de la prédiction de la difficulté des phrases dans l'ABSA et propose une méthodologie pour son évaluation. Nous proposons les réponses suivantes aux questions de recherche : **(QR1)** - "La diversité thématique et syntaxique augmente la difficulté de l'ABSA, rendant la tâche plus ardue pour les modèles traditionnels.", **(QR2)** - "Les caractéristiques linguistiques sélectionnées offrent une base pour estimer la difficulté, mais leur efficacité varie selon les contextes. D'autres caractéristiques plus sophistiquées méritent d'être explorées.", et **(QR3)** - "Les Transformeurs, notamment BERT, surpassent les modèles traditionnels dans la prédiction de la difficulté, en particulier sur des ensembles de données vastes et diversifiés, grâce à leur capacité à capturer des nuances sémantiques complexes."

Nos résultats encouragent la poursuite de la recherche sur des caractéristiques linguistiques avancées et l'application de modèles d'apprentissage profond pour améliorer la prédiction de la difficulté dans l'ABSA. Des travaux futurs pourront explorer des définitions plus nuancées de la difficulté et l'impact de la diversité linguistique et thématique sur la performance des modèles. La nécessité de corpus annotés plus vastes et diversifiés est également soulignée pour entraîner efficacement ces modèles.

## References

- [1] A.-G. Chifu, S. Fournier, Linguistic features for sentence difficulty prediction in ABSA, 2024. [arXiv:2402.03163](https://arxiv.org/abs/2402.03163).
- [2] M. Wankhade, A. C. S. Rao, C. Kulkarni, A survey on sentiment analysis methods, applications, and challenges, *Artificial Intelligence Review* 55 (2022) 5731–5780.
- [3] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, Semeval-2014 task 4: Aspect based sentiment analysis, in: the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 27–35.
- [4] F. Hamborg, K. Donnay, Newsmtsc: (multi-)target-dependent sentiment classification in news articles, in: 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021), 2021.
- [5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).