# Observation de l'expérience client dans les restaurants

Iris Eshkol-Taravella, Hyun Jung KANG
UMR 7114 MoDyCo - CNRS, Université Paris Nanterre, France
ieshkolt@parisnanterre.fr, hyunjung.kang@parisnanterre.fr

RÉSUMÉ

Ces dernières années, les recherches sur la fouille d'opinions ou l'analyse des sentiments sont menées activement dans le domaine du Traitement Automatique des Langues (TAL). De nombreuses études scientifiques portent sur l'extraction automatique des opinions positives ou négatives et de leurs cibles. Ce travail propose d'identifier automatiquement une évaluation, exprimée explicitement ou implicitement par des internautes dans le corpus d'avis tiré du Web. Six catégories d'évaluation sont proposées : opinion positive, opinion négative, opinion mixte, intention, suggestion et description. La méthode utilisée est fondée sur l'apprentissage supervisé qui tient compte des caractéristiques linguistiques de chaque catégorie retenue. L'une des difficultés que nous avons rencontrée concerne le déséquilibre entre les classes d'évaluation créées, cependant, cet obstacle a pu être surmonté dans l'apprentissage grâce aux stratégies de sur-échantillonnage et aux stratégies algorithmiques.

ABSTRACT

**Mapping Reviewers' Experience in Restaurants**

In opinion mining and sentiment analysis, studies have focused on extracting either positive or negative opinions from text and determining the targets of these opinions. Beyond the opinion polarity and its target, we propose a corpus-based model that detects evaluative language at a finer-grained level. Moreover, whereas previous works assume that classes are evenly distributed, the classes were highly imbalanced in our work. We chose a dataset of online restaurant reviews in French. We used machine learning methods to detect and classify evaluative language. Furthermore, we used resampling and algorithmic approaches to deal with class imbalance problem.

MOTS-CLÉS : Fouille d'opinion, Langage évaluative, Données disparates, Expérience client.

KEYWORDS: Opinion mining, Evaluative language, Imbalanced data, Customer experience.

## 1 Introduction

Les avis en ligne aboutissent aujourd'hui à une production abondante de données véhiculant l'évaluation du consommateur vis-à-vis de son expérience. En TAL, c'est le domaine de la fouille d'opinions qui s'occupe de cette évaluation. Pour exprimer l'évaluation, plusieurs termes peuvent être utilisés : opinion, sentiment, attitude, affect, subjectivité, etc. (Benamara *et al.* 2017). Les travaux existants se distinguent selon trois axes : la fouille d'opinion au niveau du document (Pang *et al.*, 2002 ; Turney, 2002), au niveau de la phrase (Hatzivassiloglou & Wiebe, 2000 ; Riloff & Wiebe, 2003 ; Yu & Hatzivassiloglou 2003 ; Kim & Hovy, 2004 ; Wiebe *et al.*, 2004 ; Wilson *et al.*, 2004 ; Riloff *et al.*, 2006 ; Wilson *et al.*, 2006) ou au niveau de l'aspect (Hu & Liu, 2004 ; Liu, 2015). Les recherches se limitent majoritairement aux seules catégories d'opinions positives et négatives (Pang *et al.*, 2002 ;

Turney, 2002 ; Hu & Liu, 2004 ; Kim & Hovy, 2004 ; Pontiki *et al.*, 2014, 2015, 2016). Cependant, ces dernières années, les recherches se sont réorientées vers l'extraction de suggestions (Brun & Hagege, 2013 ; Negi *et al.* 2015, 2016, 2018). Ainsi l'atelier SemEval 2019 propose des tâches visant l'extraction des suggestions dans des forums et des avis sur les hôtels. Cette tendance montre que l'intérêt que les chercheurs ont pour l'évaluation sur le web dépasse largement la notion d'opinion positive ou négative. Ainsi Benamara *et al.* (2017) proposent la notion d'intention lors d'évaluation d'un produit.

Dans cet article, nous nous intéressons aussi à la détection automatique de l'évaluation. Nous considérons que l'évaluation est une notion complexe puisqu'elle ne se limite pas juste à une valeur positive ou négative. Nous proposons un modèle d'évaluation fondé sur l'observation manuelle du corpus des avis postés en ligne sur des restaurants. Ce modèle est composé de 4 catégories : l'*opinion* (*positive*, *négative*, *mixte*), la *suggestion*, l'*intention* et la *description* [1]. Les expériences de leur détection sont fondées sur un apprentissage supervisé qui tient compte des caractéristiques linguistiques de chaque catégorie. L'une des difficultés que nous avons rencontrée concerne le déséquilibre entre les classes d'évaluation créées, cependant, cet obstacle a pu être surmonté dans l'apprentissage grâce aux stratégies de sur-échantillonnage et aux stratégies algorithmiques.

## 2 Modélisation

L'objectif du travail présenté est de détecter l'évaluation d'un restaurant exprimée soit explicitement, soit de manière implicite par un consommateur en ligne. Notre modèle s'appuie sur quatre éléments que l'on détaillera plus bas : l'opinion (positive, négative, mixte), la suggestion, l'intention et la description.

**L'opinion** : il s'agit d'une notion générale largement utilisée dans la littérature. Elle concerne l'idée que le consommateur se fait du restaurant et comment il le qualifie, le lexique évaluatif constitue donc l'indice le plus important pour détecter l'opinion. Dans le corpus, les adjectifs comme « bon », « excellent », « parfait » ou « délicieux » sont généralement associés aux opinions positives, tandis que les termes « cher », « dommage », « déception » et « bruyant » possèdent une connotation négative. Dans cette étude, les opinions telles que « pas mal », « correct », « sans plus » ont été considérées comme des opinions positives de faible intensité. La polarité d'une phrase donnée peut cependant varier selon les éléments contextuels (aussi nommés modificateurs) qui infléchissent la valeur initiale d'un terme (la négation, les intensifieurs, les atténuateurs, les conjonctions). On parle de polarité mixte lorsque l'opinion comporte les deux polarités (positive et négative). Dans de nombreux cas, celles-ci coexistent et s'articulent autour de la conjonction « mais ». Plus précisément, une polarité s'y trouve inversée, comme dans les exemples suivants : « Plat très bon, mais dessert médiocre. », « Accueil très sympathique mais cuisine décevante ». Une analyse de notre corpus montre que 64% des opinions mixtes contiennent la conjonction « mais ». En outre, certains locuteurs peuvent donner une évaluation positive en montrant leur gratitude à travers les interjections comme « bravo », « merci », etc., en mentionnant le chef ou d'autres membres du personnel : « Bravo au chef ! », « Merci pour ce bon dîner. ».

**La suggestion** : c'est une expression d'un conseil émis par un consommateur. Les suggestions peuvent être adressées à la fois aux restaurants (afin qu'ils prennent conscience des problèmes) et aux autres clients potentiels (futurs consommateurs). Elles sont souvent exprimées par les verbes

1. Les classes seront détaillées dans la section 2.

dénotant l'action. Il s'agit d'une action qui doit être réalisée par le restaurant visité ou par les futurs visiteurs selon le conseil de l'auteur d'un avis. La suggestion se manifeste fréquemment à travers le pronom « vous » ou l'impératif de la deuxième personne du pluriel, ainsi qu'à travers le conditionnel. Par exemple : « **N'hésitez** pas à venir découvrir ce restaurant, **vous** ne le **regretterez** pas » ou « Une lumière un peu plus tamisée **aurait été parfaite** ».

**L'intention** : lorsque les internautes communiquent leurs souhaits de revenir ou pas dans un restaurant, nous parlons plutôt d'intentions. Benamara *et al.* (2017) utilisent le terme *intent* (en anglais) qui recouvre entre autres les désirs, les préférences et les intentions. Selon les auteurs, les intentions sont les actions qu'un humain va entreprendre pour satisfaire ses désirs. Les intentions montrent ainsi un engagement volontaire du locuteur, une action engagée mais cette fois par le locuteur lui-même ; deux indices lexicaux retenus sont les verbes au futur et le préfixe verbal "re-" utilisé généralement pour l'itération d'une action « refaire », « revenir », « renouveler », « retourner » etc.

**La description** : si les trois éléments précédents représentent l'évaluation exprimée par un locuteur, la description concerne plutôt les informations neutres associées à l'expérience vécue, qui ont peu de rapport avec une évaluation : des éléments comme la raison pour laquelle les consommateurs se rendent dans ce restaurant, les personnes qui les accompagnent, etc. Par exemple : « Nous avons dégusté en entrée une raviole au bœuf pour les uns, tartare de saumon pour les autres. », « Un choix pour un dîner en famille, afin de fêter les 18 ans de notre dernière fille. » et « J'avais réservé pour 21 heures. ».

# 3 Méthodologie

Afin de détecter automatiquement l'évaluation modélisée à travers les 6 catégories dans le corpus d'avis, nous avons suivi plusieurs étapes (voir la Figure 1). Les avis ont été d'abord prétraités, segmentés et transformés en vecteurs. Pour l'apprentissage automatique supervisée, nous avons utilisé *Scikit-learn* (Pedregosa et al., 2011). Nous avons choisi de tester trois algorithmes de classification supervisée qui sont couramment utilisés pour ces types de tâches : *Naïve Bayes*, *Support Vector Machine (SVM)* et *Logistic Regression*. Chacun de ces algorithmes produit un modèle global qui effectue une prédiction parmi les six possibilités. Notre objectif est donc d'obtenir le meilleur score de classification.

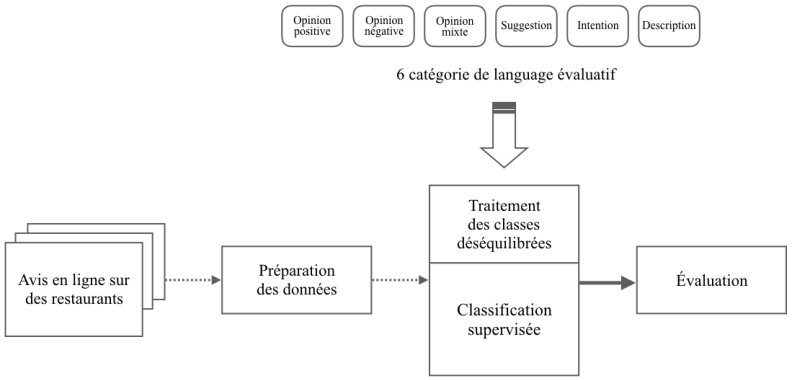


FIGURE 1: Schéma de la chaîne de traitements

## 3.1 Données

Le corpus a été collecté sur le site LaFourchette.com [2]. 200 avis par restaurant ont été extraits (au total 21 158 avis sur 126 restaurants). Le corpus a été segmenté en phrases selon les signes de ponctuation et annoté manuellement par trois annotateurs. Une des 6 catégories a été attribuée à chaque phrase segmentée. Le corpus ainsi annoté contient 2 943 phrases avec une moyenne de 10 mots par phrase. Si une phrase comprend deux catégories, notamment une opinion et une autre catégorie (intention ou suggestion), la phrase est annotée selon l'autre catégorie. Nous justifions ce choix par le fait que, mis à part les opinions, toutes les autres catégories sont peu représentées dans notre corpus. Ce problème de classes déséquilibrées est détaillé dans la section 4.3. L'accord inter-annotateur a été calculé en utilisant la mesure Kappa de Fleiss (Fleiss, 1981). Nous avons obtenu un score de 0,90, considéré comme 'presque parfait' selon l'échelle de Landis et Koch (1977). Si l'on observe la répartition de chaque catégorie (cf. Figure 2), les opinions positives représentent 68% de toutes les évaluations annotées, alors que les descriptions ou les intentions en font que 3% et 4%. Ainsi, la distribution des classes est fortement déséquilibrée. Dans la section 4.3, nous montrerons comment le problème de cette disproportion a été résolu.

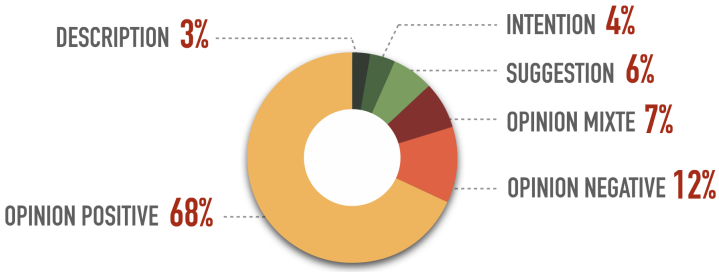


FIGURE 2: La répartition des classes dans le corpus de référence

## 3.2 Préparation des données textuelles

**Pré-traitements** : nous avons effectué le pré-traitement du corpus de la façon suivante.

- Passage des mots en minuscules ;
- Conversion des nombres par « NUM », « € » par « euros » et « % » par « pourcent » ;
- Remplacement des émoticônes par « emoPOS » ou « emoNEG » selon la polarité [3] ;
- Segmentation en phrases en suivant les signes typographiques (. ! ?) ;
- Suppression de la ponctuation. Cette étape doit être effectuée après le remplacement des émoticônes et la segmentation en phrases. Cet ordre est obligatoire pour que les émoticônes et les signes typographiques ne soient pas supprimées avant ;
- Transformation des abréviations, ce qui transforme « resto » en « restaurant » et « min » en « minutes » ;
- Lemmatisation en utilisant *TreeTagger* (Schmid 1994) ;
- Correction orthographique porte sur les mots inconnus [4] apparaissant plus que 2 fois dans le corpus.

2. La Fourchette, `https://www.lafourchette.com`
3. D'une manière générale, les émoticones sont peu présents dans notre corpus et sont toujours porteurs d'une polarité.
4. Considéres inconnus par le lemmatiseur.

**Représentation vectorielle des mots** : nous avons employé *CountVectorizer* et *TfidfVectorizer* de Scikit-learn. Nous avons testé deux paramètres dont ces méthodes disposent, *n_gram* et *max_feature*.

**Extraction de caractéristiques** :nous avons utilisé une variété de traits allant de simples traits de surface (e.g., le nombre total de caractères, de mots et la longueur moyenne de ces derniers) à des traits plus complexes comme la catégorie morphosyntaxique jugée pertinente (e.g., verbe au conditionnel, adjectif et pronom possessifs, chiffre etc.) proposée par *TreeTagger*, une valeur de la polarité et de la subjectivité identifiées par *TextBlob* [5], les mots d'opinions négatifs identifiés aussi par *TextBlob* (e.g., « déception », « désagréable », « cher », « bruyant »), la négation ainsi que la fréquence de la conjonction « mais ».

## 3.3 Traitement des classes déséquilibrées

Les travaux sur la classification de texte supposent en général une distribution équilibrée de données entre les classes (Moraes *et al.*, 2013 ; Vinodhini *et al.*, 2013 ; Zhang *et al.*, 2011 ; Ye *et al.*, 2009 ; Prabowo *et al.*, 2009 ; Tan *et al.*, 2009 ; Pang *et al.*, 2008 ; Dave *et al.*, 2003 ; Pang *et al.*, 2002). Cependant, dans la pratique, la distribution des classes est souvent asymétrique. Potts (2011) a montré que les notes des avis sur internet sont biaisées en faveur du pôle positif. Jurafsky (2014) explique ce phénomène par le principe de *Pollyanna*, une tendance générale des personnes à préférer l'information positive. La disproportion est donc un fait inévitable. À cause de ce déséquilibre, la classification est orientée en faveur de la classe majoritaire alors que les informations provenant des classes minoritaires ne sont pas prises en compte. Afin de gérer le déséquilibre entre les 6 catégories annotées, nous avons adopté différentes stratégies que l'on peut rassembler en deux groupes principaux : les stratégies d'échantillonnage et les stratégies algorithmiques (cf. le Tableau 1). Les stratégies d'échantillonnage consistent à dupliquer les observations de la classe minoritaire (*sur-échantillonnage*) ou à enlever celles de la classe dominante (*sous-échantillonnage*). Nous avons choisi le sur-échantillonnage car la classe minoritaire est sous-représentée dans les données annotées. Trois techniques de sur-échantillonnage disponibles dans *Imbalanced-learn* [6], (Lemaître *et al.* 2017) ont été utilisées : *Random Over-sampling*, *SMOTE* et *ADASYN*. Les stratégies algorithmiques consistent à pénaliser une classe sur-représentée ou à employer les méthodes ensemblistes. Il est possible de pénaliser une classe majoritaire en ajustant le paramètre *class_weight* au mode équilibré (‘*balanced*’), le poids d'une classe étant pondéré en fonction de la proportion des classes. Les méthodes ensemblistes (*Bagging*, *Boosting*) permettent également de classifier des données déséquilibrées mais elles n'ont pas été testées dans cette étude.

| **Stratégies d'échantillonnage** | **Stratégies algorithmiques** | |
|---|---|---|
| [Sur-échantillonnage] | Pénalisation | Méthodes ensemblistes |
| Random Sampling<br>SMOTE<br>ADASYN | Paramètre<br>*class_weight* | Bagging<br>Boosting<br>Simple vote |

TABLE 1: Différentes stratégies pour traiter des classes déséquilibrés

5. Une librairie Python pour traiter les données textuelles, `https://textblob.readthedocs.io/en/dev/index.html`

6. Un package Python pour traiter les données disparates (Lemaître *et al.* 2017).

## 3.4 Expérience et résultats

Pour apprendre les six étiquettes proposées, trois algorithmes de classification supervisée ont été utilisés : Naïve Bayes, Support Vector Machine (SVM) et Logistic Regression. Les expériences ont été effectuées avec une procédure de grille de recherche (*GridSearch* [7]) en utilisant une validation croisée à 5 plis. Lors des expériences, nous avons testé 10% de données. Nous y avons également associé les stratégies d'échantillonnage et d'algorithmiques (section 4.3) sur les données d'apprentissage, à l'exception de Naïve Bayes, qui ne possède pas de paramètre *class_weight*. 11 combinaisons différentes ont été testées. L'évaluation a été faite en termes de précision, rappel et F-mesure (macro moyenne). La macro F-mesure, donnant un poids identique à chaque catégorie sans tenir compte de leur taille, est donc pertinente lorsqu'il s'agit d'évaluer des données déséquilibrées (Müller & Guido, 2016). Le sur-échantillonnage de ADASYN utilisé avec l'algorithme SVM semble donner la meilleure macro F-mesure (0,79). Les résultats sont présentés dans le Tableau 2. La performance de Naïve Bayes est plus faible que les deux autres algorithmes, particulièrement sur les opinions mixtes. Ce résultat est dû au classifieur qui analyse les traits indépendamment les uns des autres. Cependant, les opinions mixtes étant composées d'opinions positives et négatives, il est difficile d'avoir des traits complètement indépendants. Nous remarquons également que la performance de la méthode classique de l'apprentissage automatique, qui ne prend pas en compte les classes disparates, est différente de celles des stratégies permettant d'avoir des classes équilibrées. Ces stratégies donnent un résultat fiable contrairement à la méthode classique qui est biaisée par la classe majoritaire. De plus, les caractéristiques de la classe minoritaire sont traitées comme du bruit et sont souvent ignorées.

| | **Macro F-mesure** | | | | |
|---|---|---|---|---|---|
| | **Classique** | **Random** | **ADASYN** | **SMOTE** | **Balanced** |
| **Naïve Bayes** | 0,66 | 0,59 | 0,61 | 0,60 | - |
| **SVM** | 0,76 | 0,77 | 0,79 | 0,78 | 0,76 |
| **Logistic Regression** | 0,78 | 0,72 | 0,72 | 0,72 | 0,77 |

TABLE 2: Macro F-mesure

Nous avons comparé nos résultats à ceux obtenus lors de l'atelier SemEval 2016 (Pontiki *et al.*, 2016). L'atelier a proposé une tâche de classification des polarités de phrases issues d'avis clients de restaurants (en anglais et en français), ce qui se rapproche de nos expériences. La tâche 5 (sous-tâche 1-3) portait sur l'attribution de la polarité (POSITIVE, NEGATIVE, NEUTRE) à une phrase donnée. La compétition utilisait cependant la métrique « Accuracy » (l'exactitude), désignant la proportion des données qui ont été classées correctement. Le meilleur accuracy (78,826) a été réalisé par Brun *et al.* (2016). La même tâche dans nos expériences a reçu l'accuracy de 87,797 en utilisant le sur-échantillonnage ADASYN associé à l'algorithme SVM. Ainsi nos résultats semblent robustes.

La précision et le rappel du sur-échantillonnage d'ADASYN utilisé avec l'algorithme SVM sont présentés dans le Tableau 3. Nous pouvons remarquer que les scores pour l'opinion positive sont élevés puisqu'ils se situent entre 0,88 et 0,96. A contrario, les autres catégories ont des résultats assez variés avec notamment une précision supérieure au rappel. Les mauvais résultats pour l'opinion mixte peuvent s'expliquer par l'implication des opinions positives et négatives dans le calcul. Autrement dit, une phrase d'opinion n'est pas toujours positive ou négative.

7. Il nous permet de trouver la meilleure combinaison d'hyper-paramètre d'un classifieur.

| | Positive | Négative | Mixte | Suggestion | Intention | Description |
|---|---|---|---|---|---|---|
| **Précision** | 0,90 | 0,86 | 0,47 | 0,94 | 1,00 | 1,00 |
| **Rappel** | 0,96 | 0,73 | 0,53 | 0,83 | 0,80 | 0,64 |

TABLE 3: Précision et Rappel

# 4 Conclusion

Les avis en ligne aboutissent aujourd'hui à une production abondante de données véhiculant l'évaluation du consommateur vis-à-vis de son expérience. L'analyse du corpus a montré qu'on ne peut pas se limiter à des notions d'opinion positive ou négative car d'autres informations et d'autres façons de les exprimer apparaissent dans le corpus. L'apprentissage automatique proposé tient compte de ces observations et des caractéristiques linguistiques de chaque catégorie retenue. Cependant, cette approche a montré un problème de déséquilibre entre les classes d'évaluation crées. Différentes approches de l'apprentissage supervisé tenant compte des spécificités du corpus d'une part et du déséquilibre de classes annotées d'autre part ont été présentées. Le sur-échantillonnage de ADASYN utilisé avec l'algorithme SVM donne la F-mesure de 0,88. En perspective, il serait intéressant de tester d'autres techniques comme des méthodes ensemblistes (*Bagging*, *Boosting*) qui pourraient donner des résultats satisfaisants. Ces méthodes combinent différents algorithmes afin d'optimiser les performances du classifieur global. Dans les challenges Netflix (2009), KDD-Cup (2009-2011) et Kaggle [8], ces méthodes ont démontré les meilleures performances (Zhou, 2012). Il serait intéressant également dans des travaux futurs de pouvoir mesurer la généricité de cette approche en appliquant sur d'autres types d'avis (messages plus longs, formes différentes, issus de corpus oraux etc.).

8. http://blog.kaggle.com/?s=1st+Place+Winner

# Références


Ashari A., Paryudi I., & Tjoa AM. (2013). Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. *International Journal of Advanced Computer Science and Applications* 4(11).

Benamara F., Taboada M. & Mathieu Y. (2017). Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics* 43(1): 201-264.

Brun, C., Perez, J. & Roux, C. (2016). XRCE at SemEval-2016 Task 5 : Feedbacked Ensemble Modelling on Syntactico-Semantic Knowledge for Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), San Diego, CA, USA.

Brun, C. & Hagege C (2013). Suggestion mining: Detecting suggestions for improvement in users' comments. *Research in Computing Science* 70.

Gopalakrishnan V. & Ramaswamy C. (2013). Performance evaluation of sentiment mining classifiers on balanced and imbalanced dataset, *International Journal of Computer Science and Business Informatics*(IJCSBI), 6 (1), 1-8.

Gopalakrishnan V. & Ramaswamy C. (2014). Sentiment Learning from Imbalanced Dataset: An Ensemble Based Method, *International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 75-87.

Gopalakrishnan V. & Ramaswamy C. (2014). Sentiment mining Using SVM-based Hybrid classification model, *Advances in Intelligent Systems and Computing*, 246, 155-162, Springer-Verlag, Berlin, Heidelberg.

Hatzivassiloglou V. & Wiebe J. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. *Proceedings of the 16th International Conference on Computational Linguistics* (COLING-2000).

Hu M. & Liu B. (2004). Mining and summarizing customer reviews. *Proceeding of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004).*

Kim SM. & Hovy E. (2004). Determining the Sentiment of Opinions. *Proceedings of the 20th International Conference on Computational Linguistics* (COLING-2004).

Jurafsky D. (2014). *The Language of Food: A Linguist Reads the Menu*. Norton.

Landis JR. & Koch GG. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Liu B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* Cambridge University Press, Cambridge.

Moraes R., Valiati JOF. & Neto WPG. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications* 40(2), 621–633.

Müller AC. & Guido S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, CA.

Negi S., Rijke M. & Buitelaar P. (2018). Open Domain Suggestion Mining: Problem Definition and Datasets. arXiv preprint arXiv:1806.02179.

Negi S., Assoja K., Mehrotra S. & Buitelaar P. (2016). A Study of Suggestions in Opinionated Texts and their automatic Detection. *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. 170–178, Berlin, Germany.

Negi S. & Buitelaar P. (2015). Towards the Extraction of Customer-to-Customer Suggestions from Reviews. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (EMNLP-2015).. 2159–2167, Lisbon, Portugal.

Pang B., Lee L. & Vaithyanathan S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceeding of 2002 conference on empirical methods in natural language*, Association for Computational Linguistics. 79–86, Philadelphia, US.

Pedregosa F., Varoquaux G., Gramfort A. *et al*. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, Journal of Machine Learning Research.

Polanyi L. & Zaenen A. (2004). Contextual valenceshifters. *Proceedings of AAAI Spring Symposiumon Exploring Attitude and Affect in Text*, 106–111.

Pontiki M., Galanis D., Papageorgiou H. *et al.* (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), San Diego, CA, USA.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). SemEval -2015 Task 1 : Aspect based sentiment analysis. *Proceedings of the 9th International Workshop on Semantic Evaluation* (SemEval-2015), 486–495, Denver, CO, USA.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). Semeval-2014 Task 4: Aspect based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation* (SemEval-2014), Dublin, Ireland.

Potts C. (2011). "On the Negativity of Negation." *Proceedings of SALT 20:* 636–59.

Riloff E., Patwardhan S. & Wiebe J. (2006). Feature Subsumption for Opinion Analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP-2006), Sydney, Australia.

Riloff E. & Wiebe J. (2003). Learning Extraction Patterns for Subjective Expressions. *Proceedings of Conference on Empirical Methods in Natural Language Processing* (EMNLP-2003), Sapporo, Japan.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Turney PD. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In P. Isabelle (Ed.), *Proceeding of association for computational linguistics 40th anniversary meeting*, ACL, 417–424. Philadelphia, PA, USA

Vásquez C. (2014). *The discourse of online consumer reviews*. London: Bloomsbury.

Wiebe J., Wilson T., Bruce R., Bell M. & Martin M. (2014). Learning Subjective Language. *Computational Linguistics*. 30(3): 277-308.

Wilson T., Wiebe J. & Hwa R. (2004). Just How Mad Are You? Finding Strong and Weak Opinion Clauses. *Proceedings of National Conference on Artificial Intelligence* (AAAI-2004).

Ye Q., Zhang Z. & Law R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36 (3), 6527-6535.

Yu H. & Hatzivassiloglou V. (2003). Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *Proceeding of Conference on Empirical Methods in Natural Language Processing* (EMNLP-2003).

Zhang Z., Ye Q., Zhang Z. & Li Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications*, 38 (6), 7674-7682.