*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4942–4947
Marseille, 11–16 May 2020


# An Empirical Examination of Online Restaurant Reviews

**Hyun Jung KANG, Iris ESHKOL-TARAVELLA**
MoDyCo UMR7114, 200 Avenue de la République, 92001 Nanterre, France
{hyunjung.kang, ieshkolt}@parisnanterre.fr

**Abstract**

In the wake of (Pang et al., 2002; Turney, 2002; Liu, 2012) inter alia, opinion mining and sentiment analysis have focused on extracting either positive or negative opinions from texts and determining the targets of these opinions. In this study, we go beyond the coarse-grained positive vs. negative opposition and propose a corpus-based scheme that detects evaluative language at a finer-grained level. We classify each sentence into one of four evaluation types based on the proposed scheme: (1) the reviewer's opinion on the restaurant (positive, negative, or mixed); (2) the reviewer's input/feedback to potential customers and restaurant owners (suggestion, advice, or warning) (3) whether the reviewer wants to return to the restaurant (intention); (4) the factual statement about the experience (description). We apply classical machine learning and deep learning methods to show the effectiveness of our scheme. We also interpret the performances that we obtained for each category by taking into account the specificities of the corpus treated.

**Keywords:** Opinion Mining, Online Restaurant Reviews, Machine Learning

## 1. Introduction

As an increasing number of people are consulting online reviews in their decision-making process, online reviews have become a precious asset in various disciplines such as marketing, information science and linguistics. Evaluation encodes the reviewers' point of view and provides subjective information based on individual preferences and tastes. With the vast amount of data in the digital age, the state-of-the-art technology alone cannot provide user's nuanced understanding. It is with a linguistic approach that we can uncover new associations between language use, social bonds, and non-linguistic aspects of communication. In this context, classifying an evaluative text as positive or negatives–as it is the case in most practical applications–is not sufficient for opinion mining (or sentiment analysis) applications. In this study, we go further to fully understand online reviews by encompassing users' needs and predicting their future actions. The main contributions of this work are:

- introducing an expanded classification scheme for evaluative language beyond the simple distinction of positive and negative;
- studying online reviews from a linguistic perspective;
- demonstrating the effectiveness of our scheme by conducting the experiments.

## 2. Related Work

While there is much literature regarding opinion mining, there are fewer works on suggestion and intention mining. We will briefly describe related works in the areas of opinion, suggestion and intention mining.

**Opinion mining.** Existing works on opinion mining are performed at document, sentence and aspect level. At the document (Pang et al., 2002; Turney, 2002) and the sentence level (Wiebe et al., 1999), the task is to detect the polarity of a given document or sentence. At the aspect level, opinions are extracted as a tuple of entity, aspect, sentiment, holder and time (Liu, 2012). (Liu, 2015) provides a good overview of opinion mining, including fundamental concepts and techniques.

**Suggestion mining.** (Ramanand et al., 2010)'s study is known to be the first attempt to extract suggestions. They approached the issue by establishing two kinds of wishes: (i) suggestions for improvements for the goods and (ii) interest in purchasing them. (Brun and Hagège, 2013) worked on product reviews and proposed a set of rules in order to detect suggestions by relying on linguistics knowledge. (Negi and Buitelaar, 2015) studied hotel and electronic product reviews in which reviewers give advice or offer suggestions to fellow consumers. (Negi et al., 2016) compared various methods of suggestion mining such as manually crafted linguistic rules, Support Vector Machines with proposed linguistic features, and deep learning approaches. At the SemEval-2019, Task 9 (Negi et al., 2019) involved suggestion mining from online reviews and forums.

**Intention mining.** Intention mining has not received much attention compared to opinion mining. (Carlos and Yalamanchi, 2012) categorized intention for the use of sales, marketing and customer service. (Chen et al., 2013) performed an intention classification to identify whether a user's post involved explicitly any intention. (Benamara et al., 2017) referred to the need for intention mining to complement sentiment analysis.

As we mentioned above, majority works have boiled down opinion mining to a problem of classifying whether a piece of text expresses positive or negative evaluation. However, evaluation is, in fact, much more complex and multifaceted, which varies depending on linguistic factors, as well as participants of the communicative activity. In (Kang and Eshkol-Taravella, 2019), we briefly introduced a classification scheme and we applied traditional machine learning methods to perform automatic classification of different evaluation categories. In order to deal with the im-

balanced class distribution, we tested two techniques (i.e., cost-sensitive method and over-sampling) and compared their performance. In this paper, we first illustrate our classification scheme for restaurant reviews from a linguistic perspective and we then describe the experiments that we conducted. We will not discuss the imbalanced class distribution issue, as it is not our primary objective of this paper.

## 3. Extended Classification Scheme

We introduce a classification scheme that detects diverse types of evaluation in a given text, in tandem with our choice of linguistic features. The scheme consists of 4 categories: opinion (positive, negative, mixed), suggestion, intention and description.

**Opinion** is the reviewer's view about different aspects of the restaurant, i.e., positive or negative assessments of aesthetic values. Below we have listed some examples of aesthetic adjectives that are inherently positive or negative.

- Positive words: *bon* 'good', *professionnel* 'professional', *chaleureux* 'warm', *délicieux* 'delicious', *joli* 'nice', *intéressant* 'interesting', *souriant* 'noisy', *confortable* 'comfortable', *efficace* 'efficient'.
- Negative words: *bruyant* 'noisy', *cher* 'expensive', *moche* 'ugly', *mauvais* 'bad', *agressif* 'agressive'.

An emotional reaction can also be attached to the evaluated restaurant as if it were some properties that the restaurant possesses. For instance, *triste* 'sad' is an adjective that concerns emotion but by saying *La salle est un peu triste et vieillotte.* 'The room is a little gloomy and old-fashioned.', it attributes a property to the restaurant of which it represents the atmosphere.
However, at numerous points, it is the combination of words that makes the polarity detection challenging. For example, valence shifters (Polanyi and Zaenen, 2006) such as negation, conjunctions, intensifiers can shift the polarity of a given sentence and result in mixed opinions. In the following sentence, *Plat très bon mais dessert médiocre.* 'Very good dish, but poor dessert.', the positive opinion is shifted to a negative one through the conjunction *mais* 'but'. Intensifiers such as *très* 'very', *trop* 'too', weaken or strengthen the original valence of a word. Furthermore, we can notice that in Computer-Mediated Communication (CMC), the intensity can be expressed in capitalization (e.g., *C'était EXCELLENT!* 'It was EXCELLENT'), emphatic punctuation (e.g., *Une soirée qui me restera en mémoire !!!* 'An evening that will be remembered!!!') and emoticons (e.g., *Merci à Adèle pour son accueil si chaleureux :-)* 'Thanks to Adèle for her warm welcome :-)').

**Suggestion** refers to the expression of advice, tips, warnings and recommendation (Negi et al., 2018), which can be addressed to both fellow consumers and business owners. Fellow consumers can get useful tips to take into account their future actions, such as *N'hésitez pas à venir découvrir ce restaurant, vous ne le regretterez pas.* 'Do not hesitate to discover this restaurant, you will not regret it'; business owners can consider suggestions to improve their business activities (e.g., *Une lumière un peu plus tamisée aurait été parfaite.* 'A more dimmed light would have been perfect.'). Suggestion, like opinion, reflects the characteristics of CMC, which combines linguistic features of writing and speaking (Georgakopoulou, 2006; Herring, 1996). For instance, reviewers occasionally speak to the readers directly and give them their suggestions, advice or warnings. The use of second-person forms (e.g., *vous* 'you', *votre* 'your'), imperatives (mostly verbs ending in '-ez'), conditional mood and explicit warnings/suggestions (e.g., *attention*, *recommender*) often indicates a suggestion.

**Intention** is "a course of action that a person or a group of persons intends to follow (Liu, 2015)". (Benamara et al., 2017) use the term intention, which includes the notion of desires, preferences and intentions. According to Benamara et al. (2017), one will take action in order to satisfy his or her desire. Thus, the intention shows a voluntary commitment of the speaker, which is initiated by himself or herself. Intention mining has the potential for practical applications. For example, studying customers' intention to purchase a product can help to identify future actions of the reviewer. In restaurant reviews, visitors evaluate the restaurant by expressing their desire to repeat (or not) the experience. In our work, we limited to explicit intention of revisiting the restaurant, as follows: *Nous y retournerons avec plaisir !* 'We'll be happy to go back!', *On reviendra !* 'We'll be back!'. We discovered two lexical indices of intention: verbs in future tense and the verbal prefix 're-'. The latter indicates repeating a previous state of being or location such as *revenir* 'to come back', *retourner* 'to return to', *refaire* 'to do again', *renouveler* 'to repeat'.

**Description.** Whereas preceding elements (i.e., opinion, suggestion and intention) represent the reviewers' evaluation of restaurants, description is more about factual information. It serves to establish shared background knowledge for the reviewer and the reader. The description typically consists of the reason for visiting the restaurant, the dishes they had, and a reference to their companions. Here are some examples: *Nous étions quatre et chacun a pris un plat différent.* 'There were four of us and each one took a different dish', *J'avais réservé pour l'anniversaire de ma maman.* 'I had made a reservation for my mom's birthday.'. The description provides practical information for customers when choosing the restaurant to visit. However, to our knowledge, this type of information is often neglected in opinion mining.

## 4. Methodology

Our objective is to detect automatically different evaluation categories (POS_OPINION, NEG_OPINION, MIX_OPINION, SUGGESTION, INTENTION, DESCRIPTION) by applying machine learning and deep learning approaches. Figure 1 shows the procedure of the experiments.

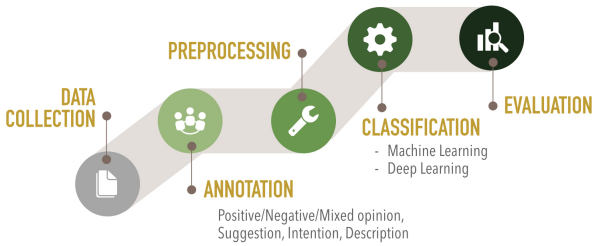


Figure 1: Procedure of the experiments

### 4.1. Data Collection

We collected online restaurant reviews written in french from the internet. Among the collected reviews, we worked on 6,287 reviews, which were segmented into sentences. As a result, the dataset consists of 17,268 sentences whose length is about 10 tokens on average.

### 4.2. Annotation

We annotated each sentence into one of six categories: POS_OPINION, NEG_OPINION, MIX_OPINION, SUGGESTION, INTENTION and DESCRIPTION. In the case where a sentence involved multiple categories, we annotated as the minority category due to the lack of its data. We evaluated the annotation task, which was done by three annotators. According to Fleiss's Kappa measure, we obtained 0.90, which is considered as 'almost perfect' (Landis and Koch, 1977). Consequently, the number of observations per category is strongly imbalanced: POS_OPINION constitutes the largest percentage at 68.2%, whereas DESCRIPTION, being the smallest, at 1.8% (see Figure 2).

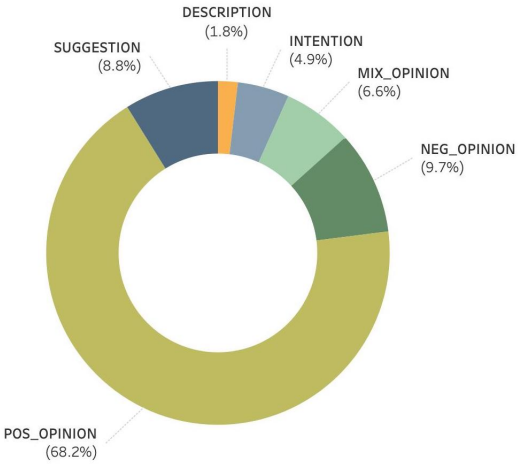


Figure 2: Class distribution of categories

### 4.3. Preprocessing

Online reviews are unstructured information and may contain various types of noise. Therefore, the process of cleaning and normalization of text is essential for the analysis. Before normalizing the text, we replaced emoticons by emoPOS or emoNEG depending on its polarity, so that they were not removed during the punctuation removal. Then we processed the following methods: lowercase conversion, punctuation removal and word normalization (handling abbreviation, replacing numbers by NUM, lemmatization using Stanford CoreNLP[1] and stemming using Snowball[2]).

### 4.4. Classification

We chose three conventional methods frequently used in text classification: linear SVM (Support Vector Machine), CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory network). We used the Python Scikit-Learn[3] library to build SVM models and Keras library[4] with the TensorFlow[5] backend for CNN and LSTM models. In Kang and Eshkol-Taravella (2019), we handled the imbalanced class distribution; however, the imbalance problem will not be discussed in this paper, as we are primarily concerned with the effectiveness of our scheme. Nevertheless, it was taken into account when evaluating the performance in Section 5.

**Linear SVM.** With classical machine learning algorithms, it is first required to manually construct the features that represent the underlying problem and then feed them in the algorithms. We tested two approaches of text representation: the Bag-of-Words (BOW) models (by using CountVectorizer and TfidfVectorizer) and the word embedding models. By employing GridSearch, a combination of CountVectorizer, unigram and bigram produced the best performance. The word embedding models are an improvement over simpler BOW models because they are capable of capturing contextual similarities between words. To develop word2vec embedding, we used Gensim[6] and applied the Continuous Bag-of-Words (CBOW) algorithm. The window size for context was set as 6, i.e., three words before and three words after the center word formed the context. Since the algorithm requires features, we crafted the following features based on an observation of the corpora: useful POS (Part-Of-Speech) tags, future tense, conditional mood, imperatives, verbs with prefix 're-', negation, emoticons, multiple punctuation, polarity and subjectivity score[7], positive and negative words[8], conjunction *mais* 'but', the euro currency ('*euro*' and €), uppercase words, length (i.e., word count) of sentence, character count, lexical diversity and lexical density. All of these features were fed into the linear SVM classifier.

[1] `https://stanfordnlp.github.io/stanfordnlp/`. Last retrieved in November 2019.

[2] `http://snowball.tartarus.org/algorithms/french/stemmer.html`. Last retrieved in November 2019.)

[3] `http://scikit-learn.org/stable/`. Last retrieved in November 2019.

[4] `https://keras.io/`. Last retrieved in November 2019.

[5] `https://www.tensorflow.org/`. Last retrieved in November 2019.

[6] `https://radimrehurek.com/gensim/`. Last retrieved in November 2019.

[7] We used Textblob. `https://textblob.readthedocs.io/en/dev/`. Last retrieved in November 2019.

[8] We used Textblob. `https://textblob.readthedocs.io/en/dev/`. Last retrieved in November 2019.

**CNN**s were generally used in computer vision; however, the idea of using them in text classification was first introduced by (Kim, 2014) whose results were promising. In order to develop a CNN model, we referred to (Zhang and Wallace, 2015), in which they propose a general configurations of hyperparameters when tuning a CNN model for text classification. The embedding layer was seeded with the word2vec embedding that was trained previously for linear SVM. The layer was followed by a one-dimension convolutional neural network, used with 16 filters and a kernel size of 3 with a ReLU (rectified linear) activation function. Subsequently, there is a one-dimension maximum pooling layer, which reduces the output of the previous layer by half. Then we have a standard flattening and a ReLU activation, followed by a softmax layer with 6 classes as output.

**LSTM.** As a variant of Recurrent Neural Network (RNN), LSTMs are capable of learning relevant context over much longer input sequences than other models. The first layer was the Embedded layer, which was the same as that of CNN. We had one LSTM layer with 100 memory units and the dropout and the recurrent_dropout were both configured as 0.2. The model then had a dense layer with an output size of 6 and a softmax activation function.

For both CNN and LSTM, we compiled the model using the categorical cross-entropy loss function and the Adam optimizer. The model was trained with a batch size of 32 with 4 epochs and 6 epochs, respectively[9]. For all experiments, we conducted a 5-fold stratified cross-validation. We produced stratified folds that contain a representative ratio of each category owing to the imbalanced class distribution.

## 5. Results

We evaluated our performance in terms of weighted-average precision, recall, F1-score and the confusion matrix. Macro-average F1-score treats each class equally, but it does not take into account the imbalanced class distribution[10]. In this regard, we considered weighted-average F1-score–which balances the class distribution–to be more appropriate for evaluating our performance. Table 1 shows the performance comparison among the classifiers. BOW+linearSVM model achieved the best performance with a weighted-average F1-score of 0.88. In this experiment, the traditional machine learning algorithms produced a better result than that of deep learning. Nevertheless, we consider that we may achieve better performance with more complex neural networks. Contrary to our expectations, the word2vec+linearSVM model had a poor performance in comparison with other classifiers. Although word2vec is said to be the silver bullet, it seems that BOW may outperform the word2vec in the case where the dataset is not large and the context is domain-specific.

[9]The number of epoch was chosen according to the results of EarlyStopping, the callback provided in Keras.

[10]`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html`. Last retrieved in November 2019.

| BOW+linearSVM | word2vec+linearSVM | CNN | LSTM |
|---|---|---|---|
| 0.88 | 0.80 | 0.85 | 0.84 |

Table 1: Weighted-average F1-score of each classifier

As shown in Figure 3, the normalized confusion matrix summarizes the classification results of the best model. Each row corresponds to the true class, while each column represents the predicted class. Among the cells on the main diagonal–which are those that were classified correctly–the DESCRIPTION's cell is slightly brighter than other classes, meaning that it is challenging to detect the category. The main reason lies in the lack of samples of DESCRIPTION and so, fewer features were trained for DESCRIPTION. Besides, everyone has different motivations, situations and stories that establish the background knowledge, which brings to a wide range of vocabularies and contexts. As a result, DESCRIPTION tends to be classified as the majority class POS_OPINION (0.41), and occasionally as NEG_OPINION (0.14). We can see a similar tendency with MIX_OPINION, whose result arises from the fact that the category involves both POS_OPINION (0.19) and NEG_OPINION (0.12) which makes it difficult to classify accurately.

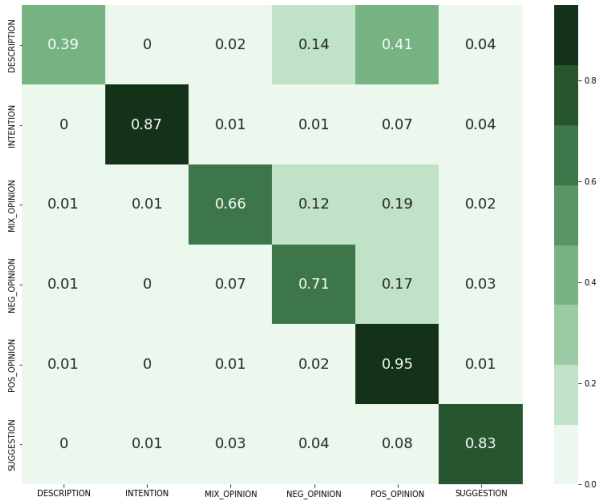


Figure 3: Confusion matrix (normalized) of BOW+linearSVM

Figure 4–6 illustrates the comparison of the weighted-average precision, recall, F1-score of among the evaluation categories and the techniques applied. In general, ML (BOW+linearSVM) had better results than DL (CNN), although the differences are small. We can observe the best performance consistently for POS_OPINION with an F1-score of around 0.94, and the second-best for INTENTION (0.86-0.88). DESCRIPTION had the worst performance with both ML (0.46) and DL (0.34), followed by MIX_OPINION with the F1-scores around 0.66. From the previous confusion matrix, we have already observed that the DESCRIPTION did not perform well compared to other categories. Moreover, we can notice throughout Figure 4–6 a significant gap in scores between ML and DL models.

With more training data on DESCRIPTION, we may get better results; however, the difficulty lies in the fact that DESCRIPTION does not appear frequently. Another solution can be oversampling, but as we have mentioned earlier, a large variety of vocabularies and contexts are used in DESCRIPTION. Therefore, producing copies (or synthetic examples) of the minority class may lead to overfitting. MIX_OPINION also seems to be tricky for classification and one of the reasons is due to the conjunction *mais* 'but', which changes the polarity of a sentence. A more fine-grained segmentation units, such as clause or phrase, can enhance the results and thus, tackling the conjunction *mais* 'but' issue can be a good starting point. For example, take the sentence that we saw earlier: *Plat très bon mais dessert médiocre.* 'Very good dish but mediocre dessert.'. Here we can divide into two statements: *Plat très bon* and *mais dessert médiocre*. Since the adversative conjunction guide the reader to more salient information, the argumentative force is combined with the latter statement. As a result, we can categorize the first statement as POS_OPINION and the second one as NEG_OPINION, and yet attribute more weights on the latter. As such, we believe that considering the conjunction for the segmentation may improve the results considerably or may even lead to the elimination of the MIX_OPINION category, given that the conjunction mais 'but' was observed in 57.5% of MIX_OPINION.

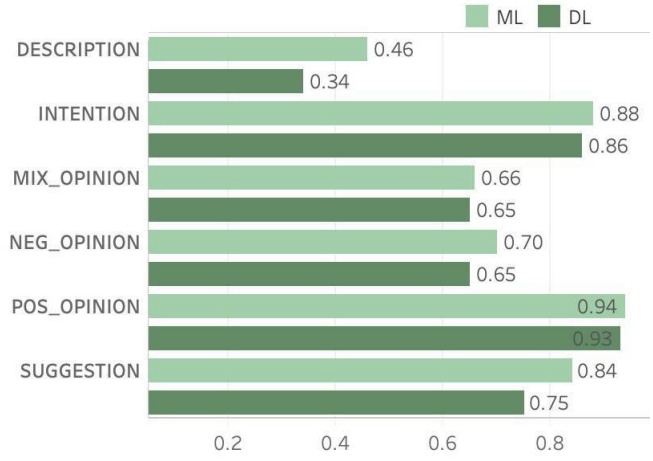


Figure 4: F1 score of BOW+linearSVC (ML) and LSTM (DL)

## 6. Conclusion

In this paper, we introduced an expanded classification scheme for evaluative language beyond the simple distinction between positive and negative. The various types of evaluative language offer meaningful insights into online reviews. We also experimentally demonstrated the effectiveness of our scheme by using ML and DL models. We discussed the results through different evaluation metrics, which helped to resolve the imbalanced class distribution problem. We obtained the best weighted-average F1-score of 0.88 with the ML model (BOW+linearSVM) and slightly behind with the CNN DL model (0.85). The proposed scheme was specific to restaurants, but it can also be applied to other places such as hotels, vacation spots, shopping malls, theaters, etc., particularly considering that getting the visitors to revisit should be one of their primary goals. Therefore, a natural extension of this work would be to study the possibility of application of our scheme in other places and different languages.

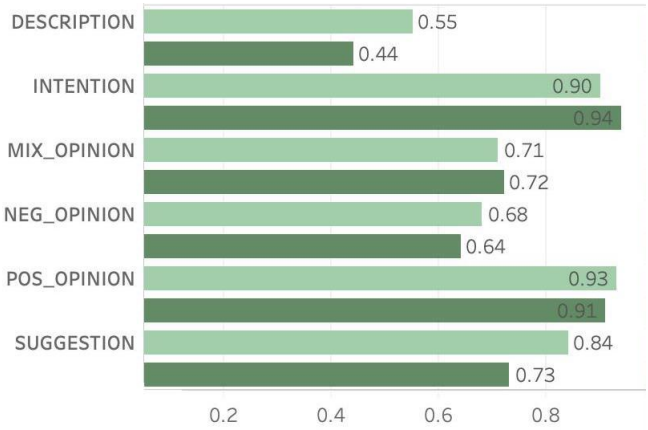


Figure 5: Precision of BOW+linearSVC (ML) and LSTM (DL)

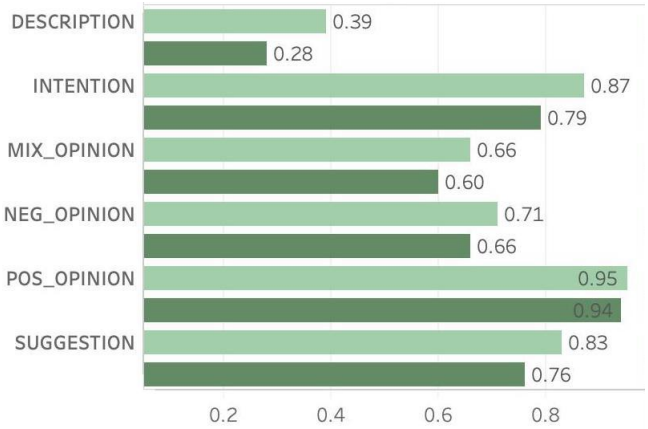


Figure 6: Recall of BOW+linearSVC (ML) and LSTM (DL)

## 7. Bibliographical References

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

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

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

Chen, Z., Liu, B., Hsu, M., Castellanos, M., and 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*,

pages 1041–1050, Atlanta, Georgia, June. Association for Computational Linguistics.

Georgakopoulou, A. (2006). Postscript: Computer-mediated communication in sociolinguistics. *Journal of Sociolinguistics - J SOCIOLING*, 10:548–557, 09.

Herring, S., (1996). *Computer-Mediated Communication: Linguistic, social, and cross-cultural perspectives*, chapter Two Variants of an electronic message schema. Pragmatics & Beyond New Series. John Benjamins Publishing Company.

Kang, H. J. and Eshkol-Taravella, I. (2019). Analysing evaluative language in online restaurant reviews. In *Proceedings of 20th International Conference on Computational Linguistics and Intelligent Text Processing*, France, La Rochelle.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

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

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

Negi, S. and 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*, 08.

Negi, S., Asooja, K., Mehrotra, S., and 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*, pages 170–178, Berlin, Germany, August. Association for Computational Linguistics.

Negi, S., de Rijke, M., and Buitelaar, P. (2018). Open domain suggestion mining: Problem definition and datasets. *CoRR*, abs/1806.02179.

Negi, S., Daudert, T., and 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)*, pages 783–883.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *EMNLP*, 10, 06.

Polanyi, L. and Zaenen, A., (2006). *Contextual Valence Shifters*, pages 1–10. Springer Netherlands, Dordrecht.

Ramanand, J., Bhavsar, K., and 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, pages 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, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wiebe, J. M., Bruce, R. F., and 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, pages 246–253, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhang, Y. and Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification.