Evaluative Language in Online Restaurant Reviews

Hyun Jung KANG, Iris ESHKOL-TARAVELLA

UMR 7114 MoDyCo - CNRS, University Paris Nanterre, France
hyunjung.kang@parisnanterre.fr, ieshkolt@parisnanterre.fr

Abstract. In the fields of *opinion mining* and *sentiment analysis*, Pang *et al.* (2002), Turney (2002) and Liu (2012, 2015), among others, have focused on extracting positive and negative opinions expressed in the text and the targets of these opinions. In contrast, beyond the opinion polarity and its target, we propose a corpus-based model that detects different evaluative language. Based on this model, we classify sentences into one of the evaluation type which is composed of four classes: (1) the reviewer's view or judgment about the restaurant (positive, negative, mixed opinion); (2) the reviewer's suggestion, advice and warning to readers, i.e., potential customers and restaurant (suggestion); (3) the reviewer's intention whether to revisit the restaurant (intention); and (4) the reviewer's neutral statement about the experience (description). Moreover, previous works assume that positive and negative classes are evenly distributed, whereas in real time application, classes are highly imbalanced (Gopalakrishnan & Ramaswamy, 2014). Similary, in our work, the number of observations per evaluation type were unequal in our work, that is 68% of positive opinions. We chose a dataset of restaurant online reviews written in French. We used, on one hand, resampling and algorithmic approaches to deal with class imbalance problem and on the other hand, supervised machine learning methods to detect and classify evaluative language. We obtained the best macro-average F1-score of 0.79 with SVM classifier and ADASYN resampling method.

Keywords: Opinion mining · Evaluative language · Imbalanced data.

1 Introduction

The number of reviews posted on the internet has exploded over the last decade. As a result, today an increasing number of people are consulting online reviews in their decision-making process. This moment of searching information from others' experiences, referred to as ZMOT (Zero Moment of Truth), has continued to grow in importance as it is becoming increasingly easy to access large amount of reviews. Among them, online review is a computer-mediated genre that we read, interact with, and even produce in our daily lives. Thus, it is important to study the language systematically that characterizes of this genre.

In this research, we study features used in evaluative language and describe how reviewers share and evaluate their experience differently in French.

Furthermore, this study makes an attempt to perform automatic classifications based on linguistic features observed in of online restaurant reviews. We will examine the effectiveness of applying machine learning techniques in order to detect different types of evaluation. Our task is different from prior works in two ways. First of all, beyond the polarity and the target of opinion–which is the objective of majority studies in opinion mining or sentiment analysis–we focus on different ways in which reviewers evaluate their experiences. Secondly, we attempt to handle the imbalance problem, owing to class disproportion in our dataset.

2 Related works

In the fields of *opinion mining* (or also called as *sentiment analysis*), *evaluative language* is used as an extended concept of opinion, sentiment, attitude, affect, subjectivity, etc. (Benamara *et al.*, 2017). For instance, Liu (2012, 2015) uses the term *opinion* to refer to evaluative language and *sentiment* to indicate positive, negative and neutral orientations. In practical application, such simplified definition is demanded. Opinion mining research has been mainly carried out at three different levels: document level, sentence level, and aspect level. The task at document-level classification is to determine whether a whole document is positive or negative (Pang *et al.*, 2002; Turney, 2002). At the sentence level, the classification can be done in two ways, either as polarity (*positive*/*negative*/*neutral*) classification or as subjectivity (*subjective*/*objective*) classification (Hatzivassiloglou & Wiebe, 2000; Riloff & Wiebe, 2003; Yu & Hatzivassiloglou 2003; Wiebe *et al.*, 2004; Wilson *et al.*, 2004; Riloff *et al.*, 2006; Wilson *et al.*, 2006). At aspect level (also called as *Aspect-Based Sentiment Analysis* (ABSA)), Liu redefines opinion as a quintuple: entity, aspect, sentiment, opinion holder, and time. *Entity* and *aspect* are together regarded as the target of opinion. *Sentiment* is implied by opinion and it can be positive, negative or neutral. *Opinion holder* is the person who conveys his or her opinion and *time* indicates the moment at which the opinion was expressed. Since each aspect of product or service can be evaluated in a decompositional manner, this approach is commonly used in practice when dealing consumer reviews. However, Benamara *et al.* (2017) argue that Liu's ABSA model does not consider the dynamic nature of evaluation, in which the interpretation of an evaluation relies on discourse and pragmatic elements. Thus, they extend the standard model by taking account for two reciprocal point of view: a linguistic perspective, in order to characterize the problem, and a computational one to compute algorithms. In this work, we considered sentence level as first step toward detecting different evaluative languages. Moreover, as argue Benamara *et al.*, we take into account linguistic contextual factors when exploring different ways reviewers evaluate their experience.

3 Conceptual Model of Evaluation in Online Restaurant Reviews

Our objective is to detect automatically evaluation, which is expressed explicitly or implicitly in a given context. To do so, we propose a conceptual model that identifies evaluative language. The model is based on four elements: *opinion* (positive, negative, mixed), *suggestion*, *intention*, and *description* (see Table 1). They will be detailed in the following sections.

Evaluation	Opinion	Polarity: positive, negative, mixed
	Suggestion	Reader: restaurant, consumer
	Intention	Polarity: positive, negative
Non-evaluation	Description	-

Table 1: Different types of evaluative language

3.1 Positive/Negative/Mixed Opinion

Opinion is the writer's view about the restaurant, i.e. what aspect of the restaurant is appreciated or not. Adjectives are the most important clue to qualify a restaurant. For instance, adjectives such as *bon* 'good', *excellent*, *parfait* 'perfect', *délicieux* 'delicious' are generally associated with positive opinions whereas *cher* 'expensive', *dommage* 'pity', *deception* 'disappointment', *bruyant* 'noisy' with negative opinions. Furthermore, positive opinions are also expressed by gratitude or encouragement to the chef and the staffs in the following manner:

- *Bravo au chef !*
  'Compliments to the chef!'
- *Merci pour ce bon dîner.*
  'Thank you for this nice dinner.'

While some adjectives are inherently positive or negative, others shift their base valance according to the context and result in mixed opinions. The most obvious *contextual valence shifters* (Polanyi & Zaenen, 2004) in our data was the connector *mais* 'but', which was observed in 64% of mixed opinions. For example, take the sentence *Accueil très sympathique mais cuisine décevante* 'The restaurant is very welcoming but the food is disappointing'. The statement *Accueil très sympathique* assess the restaurant's service positively. But then the connector *mais* is associated to the negative assessment *cuisine décevante*, it negates the positive force of the evaluation which was applied to the restaurant's service. Therefore, *mais* in this sentence reinforces the effect of the negative assessment.

3.2 Suggestion

Some reviewers provide their suggestion, advice, and warnings to readers – moving beyond conveying a personal evaluation of a product or service, to a reader-directed text. Reviewers occasionally use second-person forms (*vous*), imperatives (mostly verbs ending in '*ez*') and the conditional mood. Since via internet it is possible to reach multiple readers, reviews may address not only to fellow consumers but also to the restaurant as can be seen in the examples below.

To fellow consumers:

- *Essay**ez** le, **vous** ne ser**ez** pas déçu !*
  'Try it, you will not be disappointed!'
- ***À recommender.***
  'Recommanded.'

To the restaurant:

- *Nous **aurions** également appréciés une explication du chef.*
  'We would also have appreciated for an explanation by the chef.'
- *On **aimerait** un petit peu plus de nourriture dans l'assiette.*
  'We would like to have a little more of food on the plate.'
- *Continu**ez** !*
  'Keep going!'

3.3 Intention

The expression of the reviewer's desire to repeat the experience implies a positive evaluation. It is commonly expressed with verbs with prefix 're', meaning to repeat a previous state of being or location, such as *revenir* 'to come back', *retourner* 'to return to', *refaire* 'to do again', *renouveler* 'to repeat'. In this work, intention is limited to those that are stated explicitly. Intention is often expressed at the end of a review and functions as an overall assessment of the experience. According to Polanyi & Zaenen (2004), 'comments at the very end of a review are accorded more weight than remarks in less prominent positions'.

3.4 Description

Preceding elements (i.e. positive/negative/mixed opinion, suggestion and intention) represent the reviewers' evaluation. By contrast, description is related to the facts and the situation, i.e. why and with whom they visited the restaurant and what they ordered. For example:

- *J'avais réservé pour 21 heures.*
  'I booked for 9 p.m.'

- *190 € pour deux avec les boissons.*
  '190 € for two people with drinks.'
- *Soirée pour notre anniversaire de mariage.*
  'Evening for our wedding anniversary.'
- *Nous avons tous pris le menu du déjeuner (entrée / plat / dessert).*
  'We all took the lunch menu (starter / main course / dessert).'

4 Methodology

The basic mechanism of classification toward evaluative language using supervised machine learning algorithms is depicted in Figure 1.

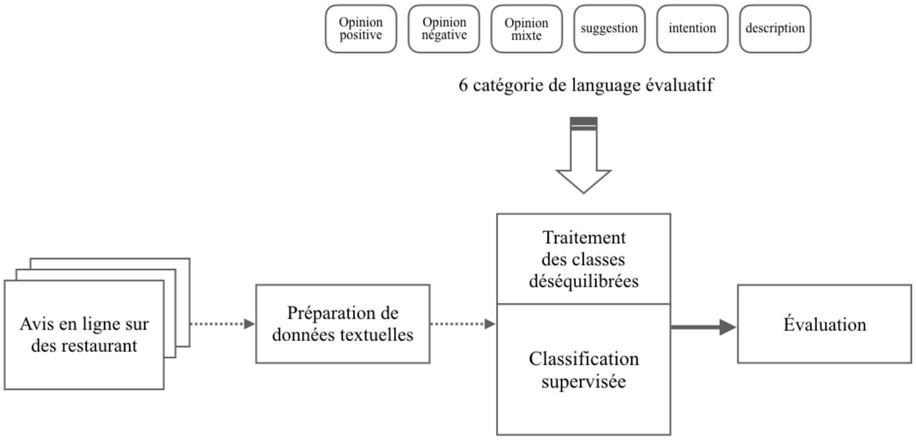
Opinion positive
Opinion négative
Opinion mixte
suggestion
intention
description
6 catégorie de language évaluatif
Avis en ligne sur des restaurant
Préparation de données textuelles
Traitement des classes déséquilibrées
Classification supervisée
Évaluation

Fig. 1: The mechanism of classification toward evaluative language using supervised machine learning algorithms

4.1 Corpus of reviews about restaurants

We have chosen Michelin-starred restaurants and non-starred restaurants, respectively 63 restaurants in Paris, from websites with restaurant reviews. The resulting dataset consists of 21,158 reviews written in French from 126 restaurants. To produce annotated data, we first segmented 1,012 reviews into sentences. Then three annotators manually annotated each sentence into one of six categories (as we have seen in Section 3): positive opinion, negative opinion, mixed opinion, suggestion, intention and description. Using *Fleiss's Kappa* measure, we obtained 0.90, which is considered 'almost perfect', according to the Landis and Koch (1977) scale. As a result, we obtained 2,943 annotated sentences, yet the class distribution is strongly unbalanced (see Figure 2). Our approaches for solving such class imbalance problems will be described in Section 4.3.

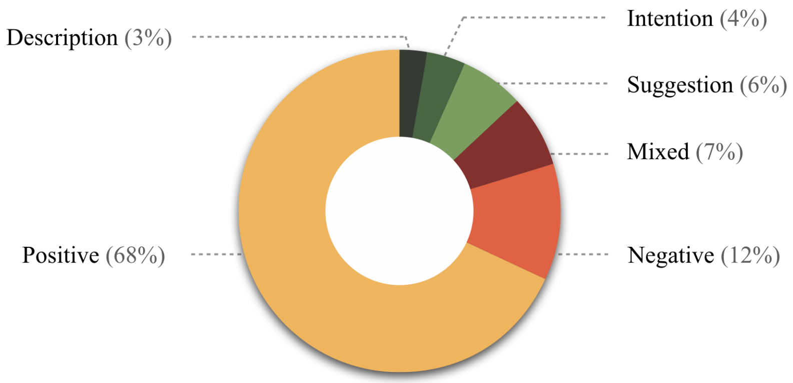

Fig. 2: Class distribution of evaluation categories

4.2 Text preprocessing and feature engineering

We applied classical pre-processing techniques such as lowercase conversion, punctuation removal, stopword removal and lemmatization, etc. to clean our textual data. While numbers are often removed in text cleaning, we kept the numbers by converting them into words because they can serve as useful information. Besides, as emoticons are widely used in internet language, they were replaced by *emoPOS* or *emoNEG* depending on its polarity. To transform our textual data into vector representations, we used `CountVectorizer,` `TfidfVectorizer` by adjusting the `ngram_range` and `max_feature` parameter which is provided in *Scikit-learn* (Pedregosa *et al.*, 2011), a machine learning library for Python. In addition, to improve the model, we created text based features as follows: word count, character count, average word density, frequency distribution of POS (Part of Speech) tags, verbs with suffix -EZ and with prefix RE-, negative word, connector *mais*, etc. Specifically, we employed *TreeTagger* (Schmid 1994) to associate each words with POS tags; then only meaningful and informative tags were filtered out such as NUM (number), VER (verb), VER:cond (verb conditional), NOM (noun), ADV (adverb), ADJ (adjective) and PRO:PER (personal pronoun). Negation involves *ne. . . pas* 'not', *ne. . . rien* 'nothing', *ne. . . jamais* 'never', *ne. . . guère* 'hardly', *ne. . . aucun(e)* 'none' and *peu* 'few'. Negative words (i.e. terms that are inherently negative) were identified by using *TextBlob*[1]. A polarity score for each word were given within the range [-1.0, 1.0] and if the score was between 0 and -1.0, we regarded the word as negative. Negative words include *déception* 'disappointment', *désagréable* 'unpleasant', *difficile* 'difficult', *excessif* 'excessive', *gênant* 'annoying', etc.

[1] A Python library for processing textual data, https://textblob.readthedocs.io/en/dev/index.html

4.3 Dealing with Imbalanced data

As shown in Figure 2, the class distribution in our data is highly imbalanced to positive opinion; this phenomenon is unavoidable. In his work, Jurafsky (2014) says that 'Reviews show humans at their most opinionated and honest, and the metaphors, emotions, and sentiment displayed in reviews are an important cue to human psychology'. Moreover, he argues that humans have a strong tendency toward the positive and optimistic (*Pollyanna principle*). It is also supported by Potts (2011) who has shown that wherever people review things on the web, the review scores are skewed toward the positive scores. For such reasons, there is always some degree of imbalance in real data sets–skewed toward the majority class, ignoring the minority classes. However researches on text classification generally assume a balanced distribution of data between the classes (Gopalakrishnan & Ramaswamy, 2014). In our work, we take into account the imbalanced distribution and we handle the problem by adopting two different approaches: data-level and algorithm-level approach (see Table 2).

Data-level	**Algorithm-level**	
(Over-sampling)	**Penalization**	**Ensemble Learning**
Random Sampling	Parameter `class_weight`	Bagging
SMOTE		Boosting
ADASYN		Simple vote

Table 2: Approaches to imbalanced data

Data-level approach. This approach consists of re-sampling the data in order to alleviate the effect caused by class imbalance. There are two techniques to make a balanced data set out of an imbalanced one: over-sampling and under-sampling. Over-sampling increases the number of minority class samples. It has the merit of not losing any information from the original dataset, as all observations from the majority and minority classes are kept. On the contrary, under-sampling reduces the number of majority samples in order to balance the class distribution. Since it removes some observations from the original data set, it might leave out important instances that provide important differences between the classes. As we do not have a large number of samples for the minority classes, we used three over-sampling techniques available in *Imbalanced-learn*[2], a package which offers a number of re-sampling techniques: *Random Over-sampling*, *SMOTE* and *ADASYN*[3].

Algorithm-level approach. One of the simplest ways is *penalization*; it is to adjust the `class_weight` parameter, which rebalances inversely proportional

[2] A python package in Python offering a number of re-sampling techniques (Lemaître *et al.*, 2017), https://imbalanced-learn.readthedocs.io/en/stable/

[3] For more details on these techniques, see Chawla *et al.* (2002) and He *et al.* (2008).

to class frequencies in the input data. On the other hand, *ensemble learning* (such as *bagging*, *boosting*, etc.) combines multiple models which can give a better performance compared to a single model. This has been the case in Netflix Competition, KDD 2009 and Kaggle competition where all of winners used ensemble learning. However, in this current work, we only used the former method.

4.4 Experiments

We carried out experiments with a mix of three supervised machine learning models that are commonly used for classification tasks (i.e., *Naïve Bayes*, *Support Vector Machine* (SVM), *Logistic Regression*) and techniques for handling imbalanced data (i.e., three over-sampling methods and penalization), which resulted in 12 different performances. In order to run supervised learning algorithms, we used Scikit-learn and to find the best combination of hyper-parameter for a given model, we applied `GridSearchCV` through 5-fold cross-validation.

4.5 Evaluating performances of the experiments

We evaluated our performance in terms of *baseline*, *precision*, *recall*, and *macro/micro average f1-score*. As no baseline has been provided for evaluative language classification, it motivated us to create a simple baseline system by using Scikit-learn's `DummyClassifier` and its `stratified strategy` parameter. This strategy makes predictions based on the training set's class distribution, which also reflects the imbalanced characteristics. The baseline takes only two common features into account: the *Bag-of-Words* (BOW) and the *Term Frequency-Inverse Document Frequency* (TF-IDF).

Macro average f1-score computes the mean of metric scores for each class, weighting each class equally. Whereas micro averaging estimates the mean for each observation, which results in minority class dominated by the majority one. If each sample is meaningful equally, we should use micro average f1-score; if we care about each class equally, it is recommended to use micro average f1-score. Since our data is highly imbalanced in class, macro average f1-score will give a sense of how effective we are on the small classes by treating each class equally. Therefore, we consider macro averaging to be more appropriate for evaluating our performance. The results in Table 3 support our choice of distinguishing macro/micro average f1-score: the micro averaging of the baseline was 0.52, while the macro averaging had poor performance (0.18). These numbers indicate that the model is better at classifying each observation sample rather than each class.The macro/micro average f1-score and the precision/recall for each class toward the baseline are shown in Table 3 and Table 4. As we can see in the tables, the results of the baseline are remarkable only with the majority class—positive opinion. This was the motivation for considering techniques regarding imbalanced data. The results and the implication of the experiment will be described in the next section.

	Precision	**Recall**	**F1-Score**
Macro average	0.19	0.17	0.18
Micro average	0.52	0.52	0.52

Table 3: Macro/micro f1-score of `"stratified"` `DummyClassifier`

	Positive	**Négative**	**Mixte**	**Suggestion**	**Intention**	**Description**
Précision	0.73	0.16	0.03	0.21	0.00	0.00
Recall	0.68	0.15	0.06	0.15	0.00	0.00

Table 4: Precision and recall of `"stratified"` `DummyClassifier`

5 Results and Implications

The macro/micro average f1-score of the experiment are indicated in Table 5. Compared to the baseline (see Table 3 and Table 4), our experiments produced better macro/micro average f1-score compared to the baseline. The results indicate that in general, SVM classifier had better value of macro-average, the best having 0.79 with ADASYN technique. Although Naive Bayes classifiers are known to outperform sophisticated classification models (Ashari, 2013), the classifier did not achieve outstanding performance in this study. Contrary to what was expected, Naive Bayes approach did not result in good performance, especially in mixed opinions class. It is due to the limitation of the classifier which assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. However a mixed opinion is composed of positive and negative opinions, which makes it difficult to have completely independent features. Moreover, micro average score is higher than macro averaging in overall. At is was the case with the baseline, the classifier has better performance in classifying each sample rather than each class.

The precision and recall are shown respectively in Table 6 and Table 7. We mostly had high precision and recall for positive opinion, which had a range bewteen 0.88 and 0.96. On the other hand, the result varied for the rest of the categories and often, with higher precision. The poor performance of mixed opinion makes sense because, as we have said above, mixed opinion involves both positive and negative opinions. Therefore, a sentence of opinion is not always positive nor negative. This is the reason why appeared the fine-grained level of study in sentiment analysis, i.e. aspect-based sentiment analysis.

		Macro average			**Micro average**		
		Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	**Random**	0.61	0.63	0.66	0.80	0.80	0.80
	ADASYN	0.62	0.65	0.61	0.79	0.79	0.79
	SMOTE	0.61	0.63	0.60	0.78	0.78	0.78
SVM	**Random**	0.84	0.72	0.77	0.87	0.87	0.87
	ADASYN	0.86	0.73	0.79	0.88	0.88	0.88
	SMOTE	0.84	0.73	0.78	0.87	0.87	0.87
	Balanced	0.82	0.72	0.76	0.87	0.87	0.87
Logistic Regression	**Random**	0.75	0.70	0.72	0.84	0.84	0.84
	ADASYN	0.73	0.71	0.72	0.85	0.85	0.85
	SMOTE	0.73	0.72	0.72	0.86	0.86	0.86
	Balanced	0.81	0.73	0.77	0.87	0.87	0.87

Table 5: Macro-average and micro-average of different methods

		Positive	**Negative**	**Mixed**	**Suggestion**	**Intention**	**Description**
Naive Bayes	**Random**	0.88	0.63	0.35	0.75	0.33	0.71
	ADASYN	0.88	0.70	0.35	0.71	0.33	0.75
	SMOTE	0.89	0.61	0.32	0.71	0.36	0.75
SVM	Random	0.91	0.81	0.43	0.88	1.00	1.00
	ADASYN	0.90	0.86	0.47	0.94	1.00	1.00
	SMOTE	0.91	0.86	0.45	0.94	1.00	0.88
	Balanced	0.88	0.79	0.44	0.88	1.00	0.88
Logistic Regression	**Random**	0.92	0.68	0.41	0.93	1.00	0.54
	ADASYN	0.92	0.72	0.45	0.94	0.80	0.54
	SMOTE	0.91	0.82	0.50	1.00	0.67	0.50
	Balanced	0.91	0.83	0.59	1.00	1.00	0.54

Table 6: Precision of different methods for each class

		Positive	**Negative**	**Mixed**	**Suggestion**	**Intention**	**Description**
Naive Bayes	**Random**	0.89	0.58	0.41	0.60	0.80	0.45
	ADASYN	0.88	0.58	0.47	0.60	0.80	0.55
	SMOTE	0.88	0.58	0.41	0.60	0.80	0.55
SVM	**Random**	0.91	0.81	0.43	0.88	1.00	1.00
	ADASYN	0.90	0.86	0.47	0.94	1.00	1.00
	SMOTE	0.91	0.86	0.45	0.94	1.00	0.88
	Balanced	0.88	0.79	0.44	0.88	1.00	0.88
Logistic Regression	**Random**	0.92	0.68	0.41	0.93	1.00	0.54
	ADASYN	0.93	0.64	0.53	0.75	0.80	0.64
	SMOTE	0.91	0.70	0.53	1.00	0.80	0.64
	Balanced	0.95	0.73	0.59	0.70	0.80	0.64

Table 7: Recall of different methods for each class

6 Conclusion

We explored various ways in which reviewers evaluate their experiences at a restaurant and we proposed a conceptual model to detect automatically evaluative language. This study applied three supervised machine learning algorithms of Naïve Bayes, SVM, and Logistic Regression and, in addition, different approaches to dealing with imbalanced data. The results of our experiments outperformed the baseline; the best macro average f1-score of 0.79 was obtained by applying SVM and ADASYN resampling. Among the categories of evaluative language, positive opinion led to good results while description and mixed opinion produced the least satisfying achievements. Further more, in order to evaluate the performance of imbalanced data, appropriate evaluation metrics were selected. Since our goal was to measure of how effective the classifier is on each type of evaluative language, we used macro averaging. A natural extension of this imbalanced data problem would be to use ensemble learning, for instance, Random Forest, XG Boost, Ada Boost, etc.

References

1. Ashari, Ahmad., Iman Paryudi, and A Min Tjoa. 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).
2. Benamara, Farah., Maite Taboada, and Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics* 43(1): 201-264.
3. Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16: 321–357.
4. Gopalakrishnan, Vinodhini., and Chandrasekaran Ramaswamy. 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.
5. Gopalakrishnan, Vinodhini., and Chandrasekaran Ramaswamy. 2014. Sentiment Learning from Imbalanced Dataset: An Ensemble Based Method, *International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 75-87.
6. Hatzivassiloglou, Vasileios and Janyce Wiebe. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. *Proceedings of International Conference on Computational Linguistics* (COLING-2000).
7. He, Haibo., Yang Bai, E. A. Garcia and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE International Joint Conference on Neural Networks* (IEEE World Congress on Computational Intelligence), pp. 1322-1328.
8. Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. *Proceeding of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD-2004).
9. Jurafsky, Dan. 2014. *The Language of Food: A Linguist Reads the Menu.* Norton.

10. Landis, J. Richard., and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
11. Lecinski, Jim. 2011. *ZMOT: Winning the Zero Moment of Truth.* Google.
12. Liu, Bing. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* Cambridge University Press, Cambridge.
13. Müller, Andreas C., and Sarah Guido. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly Media, CA.
14. Pang, Bo., Lillian Lee, and Shivakumar Vaithyanathan. 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.
15. Pedregosa, Fabian., Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, Journal of Machine Learning Research.
16. Polanyi, Livia., and Annie Zaenen. 2004. Contextual valenceshifters. *Proceedings of AAAI Spring Symposiumon Exploring Attitude and Affect in Text*, 106–111.
17. Potts, Christopher. 2011. On the Negativity of Negation. *Proceedings of SALT* 20: 636–59.
18. Riloff, Ellen, Siddharth Patwardhan, and Janyce Wiebe. 2006. Feature Subsumption for Opinion Analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP-2006).
19. Riloff, Ellen and Janice Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. *Proceedings of Conference on Empirical Methods in Natural Language Processing* (EMNLP-2003).
20. Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of of International Conference on New Methods in Language Processing.*
21. Turney, Peter D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceeding of association for computational linguistics 40th anniversary meeting*, ACL, 417–424.
22. Wiebe, Janyce, Theresa Wilson, Rebecca F. Bruce, Matthew Bell, and Melanie Martin. Learning Subjective Language. *Computational Linguistics*, 2004. 30(3): 277-308.
23. Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa. 2004. Just How Mad Are You? Finding Strong and Weak Opinion Clauses. *Proceedings of National Conference on Artificial Intelligence.*
24. Yu, Hong and Vasileios Hatzivassiloglou. 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).