

# Effective Opinion Words Extraction for Food Reviews Classification

Phuc Quang Tran<sup>1</sup>

Department of Foreign Languages  
and Informatics  
People's Police College II  
HCM City, Vietnam

Ngoan Thanh Trieu<sup>2</sup>

College of Information and  
Communication Technology  
Can Tho University  
Can Tho, Vietnam

Nguyen Vu Dao<sup>3</sup>

Department of Physical Education  
Can Tho University  
Can Tho, Vietnam

Hai Thanh Nguyen<sup>4</sup>

College of Information and Communication Technology  
Can Tho University  
Can Tho, Vietnam

Hiep Xuan Huynh<sup>5,\*</sup>

College of Information and Communication Technology  
Can Tho University  
Can Tho, Vietnam

\*Corresponding Author

**Abstract**—Opinion mining (known as sentiment analysis or emotion Artificial Intelligence) holds important roles for e-commerce and benefits to numerous business and organizations. It studies the use of natural language processing, text analysis, computational linguistics, and biometrics to provide us business valuable insights into how people feel about our product brand or service. In this study, we investigate reviews from *Amazon Fine Food Reviews* dataset including about 500,000 reviews and propose a method to transform reviews into features including Opinion Words which then can be used for reviews classification tasks by machine learning algorithms. From the obtained results, we evaluate useful Opinion Words which can be informative to identify whether the review is positive or negative.

**Keywords**—Review classification; opinion words; machine learning; important features; Amazon

## I. INTRODUCTION

Along with the strong development of Internet, e-commerce applications, social media such as reviews, forums, blogs, Facebook are increasingly popular. In order to effectively exploit the source of opinion data that users have implemented to evaluate products or raise their views on an issue they are interested in. From there, providing them with useful decisions suitable for individuals, organizations, opinion mining or sentiment analysis system is considered as a decision support tool. The main purpose of opinion mining [19][20] is the research to analyze, calculate the human viewpoints, assessments, attitudes and emotions about objects. such as products, services, organizations, individuals, problems, events, topics, and their various aspects. Opinion mining into three issues as follows: Document-based opinion mining: this is the level of simplicity of opinion mining, the document contains a point of view about a main object expressed by the author of the document. There are two main ways to explore material-based perspectives supervised and unsupervised learning. Sentence-based opinion mining: A single document can contain multiple perspectives even on similar entities. When a more detailed analysis of the various perspectives expressed in the entity documents is sought, a point-based concept mining is carried out. Aspect-based or feature-based opinion mining: research is a research

problem focused on identifying all Emotional manifestations in a given document and the aspects they refer to. The previous two methods work well when the entire document or each sentence refers to a single entity. However, in many cases, when referring to entities with many aspects (many attributes) and different views on each of the above aspects. This usually happens in product reviews or in discussion forums specific to specific product categories.

Currently, the main approaches to building a opinion mining system include lexicon-based approach [21] and approach machine learning-based [22], hybrid-based approach [22], and recently there is an in-depth approach (deep learning-based) [24]. For lexicon-based approaches, the sentiment dictionary and sentinel words are used to determine polarity. There are three techniques [20] for building an emotional vocabulary: manual-based, corpus-based, and dictionary-based approach. These methods have the advantage that emotional vocabulary has broad knowledge. However, the finite number of words in the vocabulary and the emotional score are permanently assigned to the words in the text [23]. For machine learning-based approach that uses classification techniques to conduct perspective classification, it consists of two data sets: training data set and test data set. Training sets are used to learn the different characteristics of a document, while test sets are used to test the effectiveness of the classifier. The approaches of machine learning method to classify views such as: specific probability classification are Naive Bayes, Bayesian Network, Maximum Entropy used [25]; classification based on decision trees [26]; linear classification as SVM (Support Vector Machine) [25] or Neural Network; rule-based classification. Machine-based approaches are adaptable and create models for contextual specific purposes. However, the applicability is low for new data because it requires the labeling data that can be expensive, the learning ability of machine learning models is weak, so the predictive accuracy is not high. For hybrid-based approaches [23] is a combination of machine learning and vocabulary-based approaches to improve classification performance. However, the drawback of this method is that the assessment documents have a lot of noise from words not related to the entity or aspect of the assessment) are usually

assigned a neutral point because the method does not detect any public opinion.

To enhance the predictive performance of the classification algorithm to achieve high classification efficiency, solve large-scale data problems and overcome limitations when exploring views for New classes are included in the training system during the test because some learned models are not able to handle new unknown classes. Ensemble methods overcome these limitations and aim to have a strong model by combining training decisions across categories rather than on a single classification. The ensemble methods have greater flexibility and generalization than a single classification. In this paper, we propose a method to convert assessments into features that include opinion words that can then be used to evaluate classification tasks using the ensemble learning algorithm. with the same weak learning sets such as Decision Tree Classification (dtc), Gradient Boosting Classifier (gbc), Random Forest (rf).

The rest of this paper is organized as follows. Section II presents related work. Section III presents our proposed method for classification for reviews based on the proposed set of feeling words. Section IV shows the experiments with three ensemble methods on Amazon Fine Foods reviews. The conclusion of the paper is presented in Section V.

## II. RELATED WORK

Text classification [9] is considered as the act of dividing a set of input documents into two or more categories where each document can be said to belong to one or multiple classes. Large growth of information flows and especially the explosive growth of Internet and computer network promoted growth of automated text classification. Development and advancements in computer hardware can provide enough computing power to allow automated text classification to be used in practical applications. Text classification is popularly used to handle spam emails, classify large text collections into topical categories, manage knowledge and also to help Internet search engines.

Numerous studies have proposed robust algorithms for Natural language processing. Jyoti Yadav et al. [1] proposed to use K-means algorithms as the preferred partitioning methodology. The algorithm removed the requirement of specifying the value of k in advance practically which is very difficult. This algorithmic program can obtain the best variety of cluster. Second algorithmic program cut back complexity. Some algorithmic programs use data structure that will be used to store information in every iteration which information will be employed in the next iteration. It increases the speed of clustering and cut back complexity.

In [4], the authors deployed The Bag of Words (BoW) model learns a vocabulary from all of the documents, then models every document by reckoning the number of times every word seems. The BoW model could be a simplifying illustration utilized in Natural Language Processing and information retrieval. BoW is employed in Computer Vision. In Computer Vision applications, the BoW model was applied to image classification tasks.

In [3] Zahra Nazari proposed a general definition of clustering is “organizing a bunch of objects that share similar characteristics”. The purpose of clustering is organizing data

into clusters. Such that there were high intra-cluster and low inter-cluster similarity. Hierarchical methods are commonly used for clustering in data mining problems. A hierarchical method can be subdivided as follows.

In [7] authors proposed a “New Hierarchical Clustering Algorithm” to reduce the terms some features selection technique should be used. TF-IDF technique was used which eliminates the most common terms and extracts only the most relevant terms from the corpus. Preprocessing was done by removing noisy data that can affect clustering results. Stopwords Removal and Stemming. Term Frequency-Inverse Document Frequency algorithm was used along with K-means and hierarchical algorithm

In [13], authors stated that Sentiment analysis and text summarising uses natural language processing, machine learning, text analysis, statistical and linguistic knowledge to analyze, identify and extract information from documents. The method was generally used to determine the emotions, sentiments, and summarising from large data and that information can be used to make some predictions. This work basically consisted of two machine learning methods Naive Bayes Classifier and Support Vector Machines(SVM).

Stefano Baccianella et al. released SentiWordNet 3.0 for English [8], built on the basis of WordNet 3.0. The authors built SentiWordNet 3.0 through two steps: (1) semi-supervised learning and (2) random transformation steps.

## III. CLASSIFICATION METHODS FOR REVIEW BASED ON THE PROPOSED SET OF FEELING WORDS

We proposed method contains four component as shown in Fig. 1. Reviews extracted from Amazon dataset are pre-processed into option words. The set of words then transformed by Doc2vec algorithm to become features and values for the dataset.

### A. Dataset Description

We have investigated food reviews data sets and run different ensemble learning models for review classification and useful feature extraction from tree-based decision algorithms. Food reviews data sets from amazon contains over 500,000 reviews. Each product review includes information on user, rating, and review text. The details of this data set are shown in Table I [18]. The investigated reviews are written by over 200,000 users for 74,258 products. 260 users have provided over 50 reviews. The average length of each review is about 56 words.

We only retain the content of reviews and their labels (negative or positive), and remove other information. We also remove duplicate reviews to ensure contents of review being unique. Through the pre-processing process, we have the training and test data set as input data for the machine learning classifier, shown in Table II. The contents of review, then will be processed and described in the next sections.

### B. Pre-Processing

The review texts always contain numbers, special characters, stop words, etc. Some words can appear frequently in sentences but they do not contribute the meaning of the

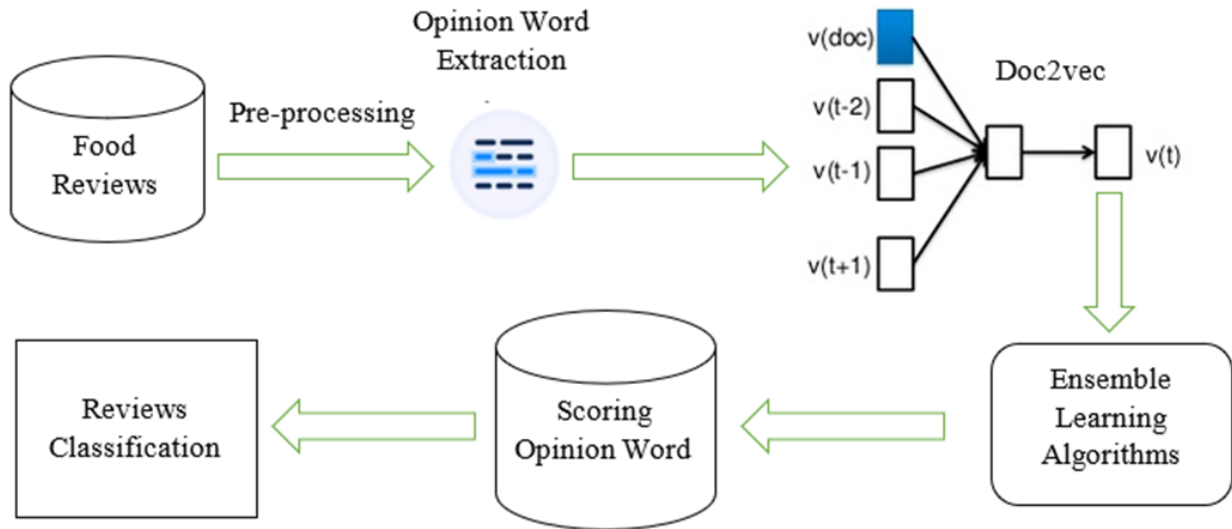


Fig. 1. An Opinion Words Extraction and Reviews Classification

TABLE I. DATA DETAIL

Dataset statistics	Number of records
Reviews	568,454
Users	256,059
Products	74,258
Users with >50 reviews	260
Median no. of words per review	56

TABLE II. DATA TRAIN AND TEST

Dataset statistics	Number of records
Reviews	18532
Train	14825
Test	3707

sentences (For example, prepositions such as on, the, at, etc.). We have separated the words in the reviews into tokens and removed the unnecessary words by using the NLTK tool [12]. This is a tool for text analysis in natural language processing. The remaining words can bring the meaning of the sentences including nouns, noun phrases, verbs, adjectives, and adverbs, so we can find participation of those types of words in each product review to determine the meaning of the reviews.

### C. Opinion Word Extraction

We performed opinion words extraction to sentiment expression from product reviews. For example such as “good”, “bad”, “great”, “better”, etc. These words are often of the adjectives, adverbs and adjectives by using the OpinionFinder [14] tool. This tool is a document processing system and automatically identifies subjective sentences as well as various aspects of subjective sentences, opinion words extraction to

sentiment expression in reviews [15]. For each product review, we have extracted opinion word expressed as the attributes of the review. Next step, we perform training on the doc2vec model to determine the weight of each opinion word and each reviews.

### D. Doc2vec

Doc2vec or Paragraph vector [11] is a method of vectorizing text. The Doc2vec method is similar to word2vec [16] but instead of representing the word vector while the Doc2vec method will represent the document as vector. There are two ways of building a Doc2vec model: Distributed Memory version of Paragraph Vector (PV-DM) and Distributed Bag of Words version of Paragraph Vector (PV-DBOW). The PV-DM model is an extension of the CBOW model of word2vec. The PV-DM model works on remembering what is missing in context, in a topic, or in a paragraph. The PV-DM model uses surrounding vectors (context) to predict target words. In addition, this model adds a vector attribute of a document. After training the word vector, the text vector will also follow. After training the word vector, the document vector will also follow. The output, the PV-DM model will give the vector index of the document. The word vector will represent the conceptual representation of a word, the document vector will represent a document.

The PV-DBOW model is similar to Skip-gram model of word2vec. But it trains faster and takes less memory than Skip-gram because there is no need to remember the words[11]. This technique is used to train models when building a set of documentation requirements. Then, each word vector is created for each word and each document vector will be created for each document. The model outputs are vectors corresponding to the calculated new input.

In this paper, we approach both the PV-CBOW and PV-DBOW models to build a set of food assessment document

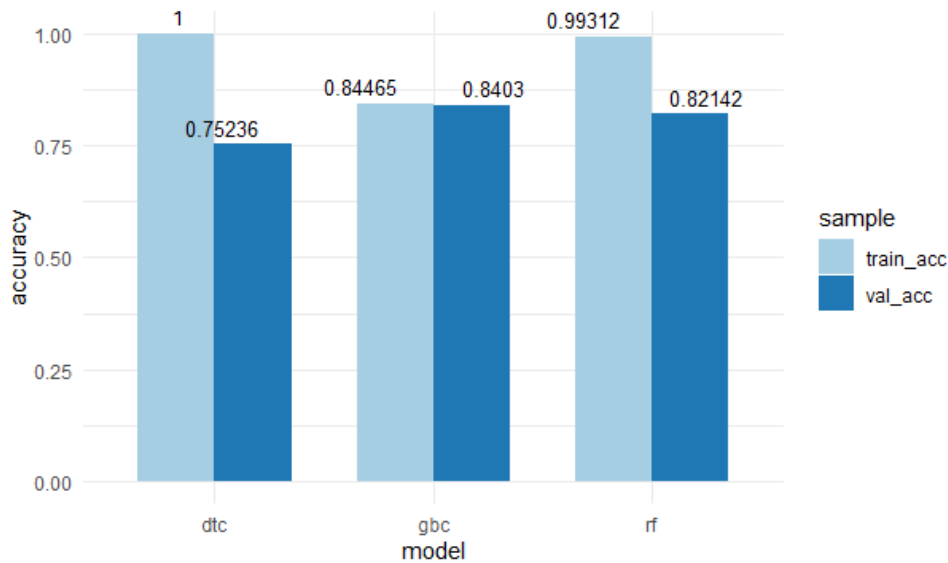


Fig. 2. Compare Accuracy of Three Classifiers DTC, RF, and GBC.

vectors based on the doc2vec model approach in the Gensim library.

#### E. Sentiment Classification

In order to classify reviews, we use robust machine learning including Gradient Boosting Classification, Random Forests and Decision Tree. Food assessment document vectors are input to classified as “positive” and “negative” by population methods with the same basic classification such as Decision Tree [2], Random Forest Algorithms [6], Gradient Boosting Classifiers [10], [5].

Decision Trees (DTs) [2] are a non-parametric supervised learning method used for classification and regression. The goal is to generate a learning model that predicts the value of a target category by learning simple decision rules inferred from the data features.

Random forests [6] or random decision forests are an ensemble learning method for classification. The algorithm is also used for regression and other tasks. This algorithm is rather robust to develop computation framework based machine learning with fast speed and give reasonable results.

Boosting [5] is a technique to transform weak learners to strong learners. Each new tree is a fit on a modified version of the original data set. Gradient Boosting Classifiers are ensemble methods with similar weak base classifiers to improve the formation of a strong learning model. The idea of the ‘Gradient Boosting’ classification is based on PAC (Probability Approximately Correct Learning) [17]. In this study, we approach the Gradient boosting classification model on the dataset including the food evaluation document vector already developed above. The gradient boosting trees can be strong method to do prediction tasks.

## IV. EXPERIMENTAL RESULTS

In this section, we evaluate the efficiency of the set of opinion words on review classification tasks with the algorithms

of Gradient Boosting Trees (GBC), Random Forests (RF) and Decision Tree (DTC).

#### A. Reviews Classification with Different Algorithms

Fig. 2 shows the prediction performance of the considered algorithms. We can see GBC reveals less overfitting compared to other algorithms. For Decision Tree Classifiers, although the training can reach to 100%, it shows the worst result in testing performance with only 0.752 in Accuracy. Otherwise, GBC only reveals the lowest performance in training phase but this algorithm can obtain to 0.845 in testing phase. Another case, Random Forest exhibits rather high both in training and testing phases. Random Forest obtain an accuracy of 0.821 which is higher than Decision Tree Classifiers but lower than Gradient Boosting Trees.

From the results obtained, we expect Gradient Boosting Trees can extract meaningful words to evaluate the reviews.

#### B. Useful Words to Determine the Meaning of the Reviews

For evaluating which words are useful to determine whether a review is negative or positive, we compute important scores extracted from all considered classifiers. As exhibited in Fig. 3, 4 and 5, words such as “good”, “great” are the most influence words on characteristics of reviews. As Observed from these figures, the word of “great” hold an important to evaluate whether the review is negative or positive. We can find easily that the statements contain “great” expressing a satisfaction on the product. “good” and “better” also convey good feelings on the product. From the 4th important word, there are some slight differences among the considered algorithms. While GBC and RF provide “happy” and “love”, respectively, which sound reasonably, the DTC shows the word of “healthy”. Although, they both bring positive feelings but “healthy” affects a narrow scope comparing to “happy” and “love”. This can lead to RF and GBC reveal better results than DTC.

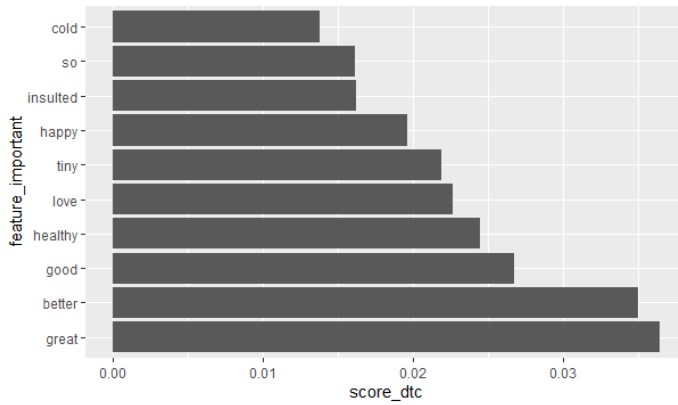


Fig. 3. TOP TEN Important Features in Food Reviews Generated from Decision Tree Classifiers.

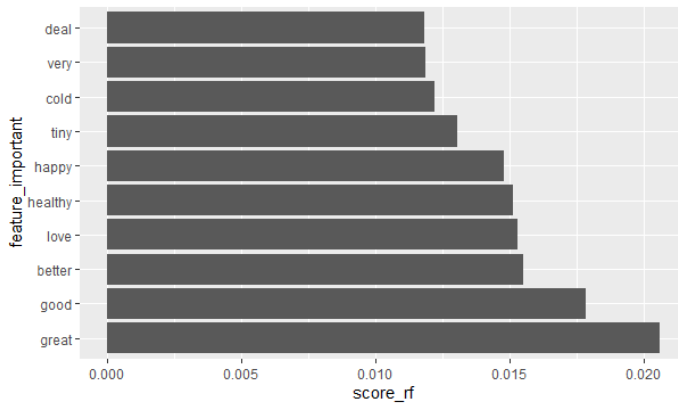


Fig. 4. TOP TEN Important Features in Food Reviews Extracted by Random Forest Algorithm.

Comparing the important scores among features shown those figures, we can see there are significant differences among the high important words in Gradient Boosting Trees while some words in other algorithms seem to be grouped such as “better” and “great” in Decision Trees or a group of “happy”, “healthy”, “love” and “better” in Random Forest.

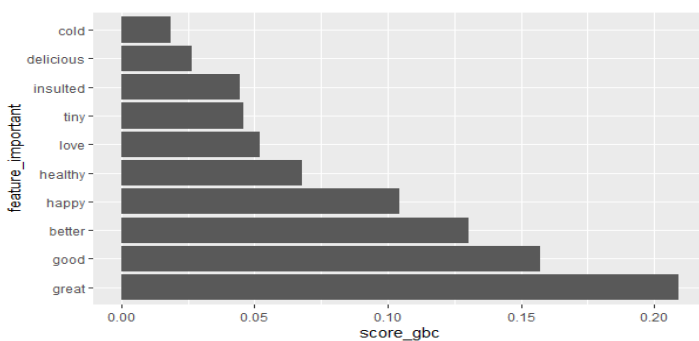


Fig. 5. TOP TEN Important Features in Food Reviews Generated from Gradient Boosting Classifiers.

## V. CONCLUSION

We introduced a set of opinion words and useful words extracted from learning model to evaluate whether a review for food production on Amazon is negative or positive. The proposed approach can be possible to apply to other e-commerce systems. The proposed features can reach a promising result with classic machine learning algorithms.

Useful words extracted from Gradient Boosting classifier and Random Forests are interesting to predict which review is positive. Some words as observed from the results are rather familiar and common words to express the feeling when we use some product.

With achievements of deep learning techniques, further research can leverage to propose sophisticated models to improve the prediction.

## REFERENCES

- [1] Jyoti Yadav, Monika Sharma. “A Review of K-mean Algorithm” International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013.
- [2] J. R. Quinlan. Induction of Decision Trees. Machine Language journal. 1986
- [3] Zahra Nazari, Dongshik Kang and M.Reza Asharif “A New Hierarchical Clustering Algorithm” ICIBMS 2015, Track2: Artificial Intelligence, Robotics, and Human-Computer Interaction, Okinawa, Japan .
- [4] Deepu S, Pethuru Raj, and S.Rajaraajeswari. “A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction” International Journal of Advanced Networking and Applications (IJANA).
- [5] Ke, Guolin et al. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” NIPS (2017).
- [6] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>.
- [7] Prafulla Bafna, Dhanya Pramod and Anagha Vaidya “Document Clustering: TF-IDF approach” International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) - 2016 University”
- [8] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. ”SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining,” in LREC 7th Conference on Language Resources and Evaluation, Valletta, MT, 2010.
- [9] Wibowo, W. and Williams, H.E. (2002). Simple and accurate feature selection for hierarchical categorisation. In Proceedings of the 2002 ACM symposium on Document engineering (pp. pp. 111-118). : ACM Press, McLean, Virginia, USA
- [10] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), pp.1189-1232, 2001.
- [11] Quoc Le, Tomas Mikolov. Distributed representations of sentences and documents. Proceedings of the 31st International Conference on International Conference on Machine Learning, Volume 32, pp.1188-1196, 2014.
- [12] Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O’Reilly Media, 2009.
- [13] Kim, S. M., and Hovy, E. (2004, August). Determining the sentiment of opinions. In Proceedings of the international conference on Computational Linguistics (p.1367). Association for Computational Linguistics.
- [14] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis. Proceedings of Interactive Demonstrations, 2005.
- [15] Lingjia Deng, and Janyce Wiebe. MPQA 3.0: An Entity/Event-Level Sentiment Corpus. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.1323–1328, 2015.

- [16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems, volumn 2, pp.3111–3119, 2013.
- [17] Leslie Valiant. Probably Approximately Correct: Nature’s Algorithms for Learning and Prospering in a Complex World. Basic Books, Inc., 2013.
- [18] Julian John McAuley, and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. Proceedings of the 22nd international conference on World Wide Web, pp. 897–908, 2013.
- [19] Bing Liu, and Lei Zhang, “A Survey of Opinion Mining and Sentiment Analysis” (Book chapter), Mining Text Data, pp.415-463, 2012.
- [20] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, 2012.
- [21] Hiroshi Kanayama, Tetsuya Nasukawa, “Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis”, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP’06), pp.355–363, 2006.
- [22] Ronen Feldman, “Techniques and applications for Sentiment Analysis”, Communications of the ACM, 56(4), pp. 82-89, 2013.
- [23] Alessia Andrea, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo, “Approaches, Tools and Applications for Sentiment Analysis Implementation”, International Journal of Computer Applications 125(3), pp.26-33, 2015.
- [24] Lei Zhang, Shuai Wang, and Bing Liu, ”Deep Learning for Sentiment Analysis: A Survey”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), pp.1-25, 2018.
- [25] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, “Thumbs up? Sentiment Classification Using Machine Learning Techniques”, Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing -EMNLP’02 (volume 10), pp.79–86, 2002.
- [26] John Ross Quinlan, “Induction of Decision Trees”, Machine Learning (volume 1), pp. 81-106, 1986.