



# Ensemble Learning for Mining Opinions on Food Reviews

Phuc Quang Tran<sup>1</sup>, Hai Thanh Nguyen<sup>2</sup>, Hanh My Thi Le<sup>3</sup>,  
and Hiep Xuan Huynh<sup>2</sup>

<sup>1</sup> Faculty of Foreign Language and Informatics,  
People's Police College II, Ho Chi Minh City, Vietnam

<sup>2</sup> College of Information and Communication Technology,  
Can Tho University, Can Tho City, Vietnam  
{nthai.cit,hxhiep}@ctu.edu.vn

<sup>3</sup> Faculty of Information Technology, University of Science and Technology  
Da Nang University, Da Nang City, Vietnam  
1tmhanh@dut.udn.vn

**Abstract.** This paper proposes an ensemble learning model for opinion mining on food reviews. The proposed model is built on an ensemble of decision trees called Random classification forest. This model performs the task of classifying sentiment about food as positive, negative, or neutral. The ensemble learning model was evaluated on two scenarios, which we built based on important features of the reviews. The experimental results on the food reviews data set have shown the effectiveness of the proposed model.

**Keywords:** Opinion mining · Opinion ensemble learning · Food reviews

## 1 Introduction

Opinion mining or sentiment analysis is widely applied in social life [1]. The aims of opinion mining to extract opinions and sentiments from natural language text using computational methods. The computational methods of opinion mining are based on the ones used in data mining, machine learning, and others [2]. Currently, there are many studies on opinion mining with supervised machine learning approaches to classify sentiment such as Naive bayes classification, Maximum entropy classification, Support vector machines, and classification based on Decision tree are used [3–5]. Ensemble methods [6,7] for sentiment classification also used with the classifiers of individual same(Decision trees) e.g. Random forest algorithm [8] for sentiment classification with data from Twitter. The accuracy of measurements in this study is around 75%. Different supervise machine learning algorithms such as Random forest, Gradient boosting,

and Voting classifier are used in sentiment classification of Music Lyrics reviews [9]. Sentiment classification results of Random forest algorithm is better than other algorithm with accuracy of 89.1%. In the sentiment analysis of reviews on social networking like WhatsApp, Facebook, and Twitter [10]. A technique used that is machine learning algorithm-Random forest to sentiment classification of comments are positive or negative. A study on sentiment analysis of Online Food Reviews using Big Data Analytics [11]. The techniques of sentiment analysis using apache spark data processing system for big datasets of Amazon Fine Food reviews. The accuracy named as Linear SVC, Logistic Regression, and Naïve Bayes have more than 80%. Moreover, sentiment analysis on food reviews was performed using machine learning algorithms such as Random forest, Naïve Bayes, SVM, and Logistic regression to classify sentiments based on review text features for multiclass classification. The resulting accuracy of the corresponding classifiers are Random forest 48.80%, Naïve bayes 42.45%, SVM 51.80%, and Logistic regression 55.40% [12].

At present, opinion mining techniques on food reviews have not yet focused on the majority sentiment classification and the important features of the reviews such as product aspects e.g. “quality”, “flavor”, “size” of the product, and sentiment orientation of opinion to opinion target of more detail, more clearly. The studies on food reviews also did not take into the holder and the time offers an opinion of the product reviews. Although this is a very important factor to sentiment orientation for opinion.

In this paper, we propose a new approach to ensemble learning for mining opinions on food reviews. We performed opinion discovery tasks on food reviews to convert the reviews into important features which are then used for classification using an ensemble learning algorithm such as Random classification forest into sentiment orientation is positive, negative, or neutral.

The rest of this paper is organized as follows. Section 2 some concepts related to opinion. Section 3 performs opinion discovery tasks. Section 4 how to build an opinion ensemble learning from features. Section 5 is an opinion summary from the previous tasks. Section 6 presents the result of the experiment, and the section finally is the conclusion of the paper.

## 2 Opinion Modeling

### 2.1 Opinion

An opinion [2] is a quadruple.

$$(g, s, h, t)$$

where  $g$  is the sentiment target,  $s$  is the sentiment of the opinion about the target  $g$ ,  $h$  is the opinion holder, and  $t$  is the time when the opinion is expressed.

### 2.2 Simplify Definition

An opinion is a quintuple [2]

$$(e, a, s, h, t)$$

where  $e$  is the target entity,  $a$  is the target aspect of entity  $e$  on which the opinion has been given,  $s$  is the sentiment of the opinion on aspect  $a$  of entity  $e$ ,  $h$  is the opinion holder, and  $t$  is the opinion posting time;  $s$  can be positive, negative, or neutral, or a rating (e.g., 1–5 stars).

In the case opinion directed at a whole entity, the special aspect GENERAL is used to express that opinion.

### 2.3 Reason and Qualifier

*Reason.* [2] A reason is the cause or explanation of the opinion.

*Qualifier.* [2] A qualifier of an opinion limits or modifies the meaning of the opinion.

### 2.4 Entity

An entity [1]  $e$  is represented by itself as a whole and a finite set of aspects.

$$A = \{ a_1, a_2, \dots, a_n \}$$

$e$  can be expressed in text with any one of a finite set of its entity expressions  $\{ ee_1, ee_2, \dots, ee_s \}$ . Each aspect  $a \in A$  of entity  $e$  can be expressed with any one of a finite set of its aspect expressions  $\{ ae_1, ae_{i2}, \dots, ae_m \}$ .

*Aspect.* The aspects  $a \in A$  of an entity  $e$  are the components and attributes of  $e$ .

### 2.5 Document

An opinion document [1]  $D$  contains opinions on a set of entities  $\{ e_1, e_2, \dots, e_r \}$  and a subset of their aspects from a set of opinion holders  $h$  contains opinions on a set of entities  $\{ h_1, h_2, \dots, h_p \}$  at some particular time point  $t$ .

## 3 Opinion Discovery

Given an opinion documents  $D$ , mining opinions consists of the following eight main tasks [2]:

*Task 1 (Entity Extraction and Resolution).* Perform the task of extracting entities expressions in documents  $D$  and group entities synonyms into clusters (or categories). Each entity represents a clustering expression of the entity.

*Task 2 (Aspect Extraction and Resolution).* Perform the task of extracting aspects expressions in document  $D$  and group aspects synonyms into clusters (or categories). Each aspect represents a clustering expression of the aspect.

*Task 3 (Opinion Holder Extraction and Resolution).* Extract the expression of the holder of each opinion from the review or structured data and group them.

*Task 4 (Time Extraction and Standardization).* Extract the posting time of each opinion and time format.

*Task 5 (Aspect Sentiment Classification or Regression).* In the case of sentiment classification, determine the aspect (or entity) whose opinion is positive, negative, or neutral. In the case of regression, determine the numeric sentiment rating score of the aspect(or entity).

*Task 6 (Opinion Quintuple Generation).* This task is generate all opinion quintuples  $(e, a, s, h, t)$  expressed in  $D$  from previous tasks.

*Task 7 (Opinion Reason Extraction and Resolution).* Perform the task of extracting reason expressions for each opinion and group reason synonyms into clusters. Each reason represents a clustering expression of the reason for the opinion.

*Task 8 (Opinion Qualifier Extraction and Resolution).* Perform the task of extracting qualifier expressions for each opinion and group qualifier synonyms into clusters. Each qualifier represents a clustering expression of the qualifier for the opinion.

For example about a review such as:

(1) These Chips are quite tasty and the price is right. (2) Packaged very well, would buy again. (Posted by: Swannie, time: Oct-4-2011).

The task should six generate the following opinion quintuples:

(Chips, GENERAL, Positive, Swannie, Oct-4-2011).  
 (Chips, Price, Positive, Swannie, Oct-4-2011).  
 (Chips, Packaged, Positive, Swannie, Oct-4-2011).

The task seven and eight are find the reasons and qualifier for opinion:

(Chips, Price, Positive, Swannie, Oct-4-2011).

The reason for opinion: none.

The qualifier of opinion: none.

## 4 Opinion Ensemble Learning

Given a set of  $n$  individual classifiers  $O = \{o_1, o_2, \dots, o_n\}$ , each  $o_i$  is an opinion quintuples  $(e, a, s, h, t)$  and  $K$  classes, each one labelled as  $c_k$  with  $k \in [1, \dots, K]$ . Given a classifier  $o_i$ ,  $i \in [1, \dots, n]$ . The vector  $K$ -dimensional  $v = (o_i^1, o_i^2, \dots, o_i^K)$  is represented for the output.

In the case of class label where  $o_i^k \in \{0, 1\}$  that obtain value 1 if the class is  $k$  and 0 otherwise.

In the case of class probability where  $o_i^k \in [0, 1]$ . An estimate of the posterior probability for classifier  $o_i$  is  $P(c_k|v)$ .

The model aggregates the classification by obtaining the value that occurs most often for each classification

$$H(v) = c_{\arg \max_j \sum_{i=1}^n h_i^j(v)}$$

where  $H(v)$  is output of opinion ensemble learning Fig. 1.

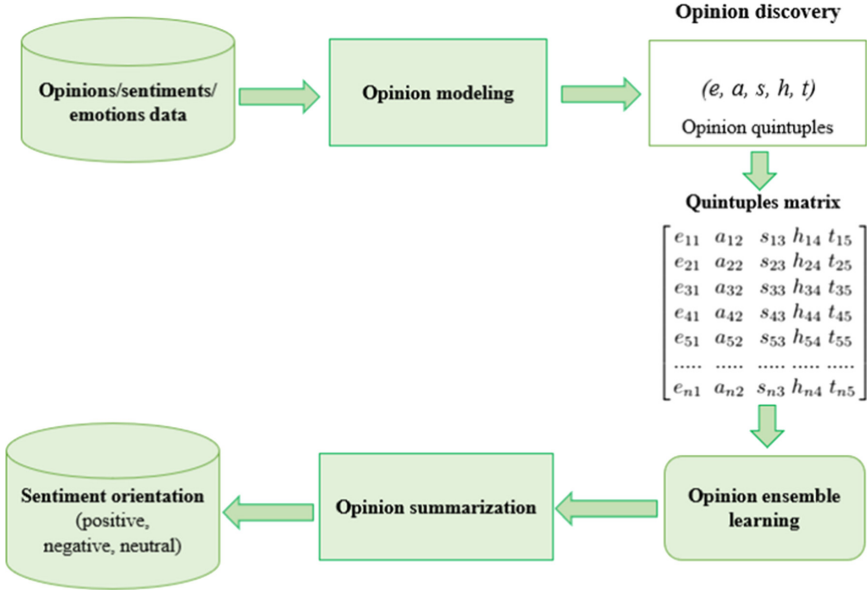


Fig. 1. Illustrate an opinion ensemble learning graph.

### 4.1 Quintuples Matrix

Given an opinion quintuple  $(e, a, s, h, t)$ . A matrix  $n \times 5$  of opinion quintuple is shown as follows:

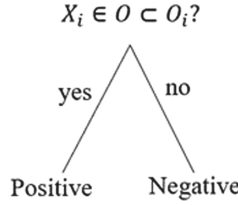
$$\begin{bmatrix} e_{11} & a_{12} & s_{13} & h_{14} & t_{15} \\ e_{21} & a_{22} & s_{23} & h_{24} & t_{25} \\ e_{31} & a_{32} & s_{33} & h_{34} & t_{35} \\ e_{41} & a_{42} & s_{43} & h_{44} & t_{45} \\ e_{51} & a_{52} & s_{53} & h_{54} & t_{55} \\ \dots & \dots & \dots & \dots & \dots \\ e_{n1} & a_{n2} & s_{n3} & h_{n4} & t_{n5} \end{bmatrix}$$

The quintuples matrix is called a feature matrix as an input to the process of building an opinion forest as the basis for mining and summarizing opinion.

### 4.2 Decision Tree

Given  $X_i$  is the sentiment orientation classification predictor variable [13] that takes values from a finite set of categories  $O_i = \{e_{i,1}, a_{i,2}, s_{i,3}, h_{i,4}, t_{i,5}\}$ . A split sends a subset of these categories  $O \in O_i$  to the left and the remaining categories to the right in Figure.2.

The tree is split one node into two nodes by looking at the split on every predictor variable and choosing the best split to be the splitting criterion.



**Fig. 2.** The split on a opinion quintuples predictor variable  $X_i$ .

In the case of classification [14] where there are  $K$  classes denoted  $1, \dots, K$  a typical splitting criterion is the Shannon entropy, given by

$$H(S) = - \sum_{k \in K} p(k) \log(p(k)) \tag{1}$$

where  $S$  is a collection of training opinion quintuples and denotes  $k$  is the class label. The collection of all classes is denoted  $K$  and  $p(k)$  designates the empirical distribution extracted from the training opinion quintuples within collection  $S$ . A measure of node purity by measuring the information gain by the following formula:

$$I = H(S) - \sum_{i \in \{L,R\}} \frac{|S^i|}{|S|} H(S^i) \tag{2}$$

The decision tree includes decision rules. The path from the root node to the leaf node is represented as root  $\rightarrow$  leaf. As a rule, to the right of the denote “ $\rightarrow$ ” has a value and to the left of the denote “ $\rightarrow$ ” has at least one value. The denote “ $\cap$ ” is the intersection mark. For example in Fig. 3 has a decision rules:

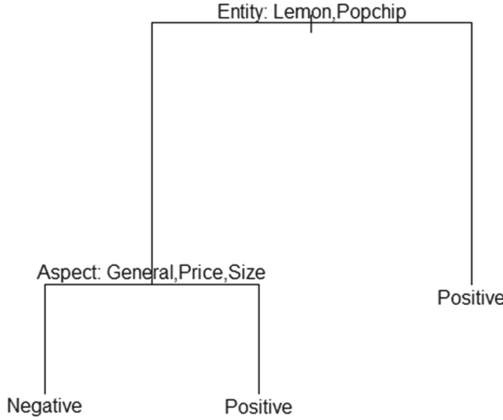
$$(Entity = Lemon) \cap (Aspect = Price) \rightarrow Opinion = Negative$$

### 4.3 Random Forests

Given a random vector have  $p$ -dimensional, denoted by  $X = (X_1, X_2, \dots, X_p)^T$  where  $X_i \in \{e_{i,1}, a_{i,2}, s_{i,3}, h_{i,4}, t_{i,5}\}$  representing the predictor variables and a random variable  $Y$  representing the response variables. Assume an unknown joint distribution  $P_{XY}(X, Y)$ . The object is to find a prediction function  $f(X)$  for predicting  $Y$ . The prediction function [15] is determined by a loss function  $L(Y, f(X))$ . This function is determined by minimizing the expected value of the loss:

$$E_{X,Y}(L(Y, f(X))) \tag{3}$$

where the subscripts denote expectation with respect to the joint distribution of  $X$  and  $Y$ ;  $L(Y, f(X))$  is a measure of how close  $f(X)$  is to  $Y$ . The loss function for a typically chosen classifier is as follows:



**Fig. 3.** An example of a decision tree for leaf node represents the sentiment.

$$L(Y, f(X)) = I(Y \neq f(X)) = \begin{cases} 0 & \text{if } L(Y, f(X)) \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

In the case of classification, if the collection of possible values of  $Y$  is denoted by  $\mathcal{Y}$ ,  $E_{X,Y}(L(Y, f(X)))$  is minimized for loss gives:

$$f(x) = \arg \max_{y \in \mathcal{Y}} P(Y = y | X = x) \tag{5}$$

Ensembles of base learners  $h_1(x), h_2(x), \dots, h_j(x)$  and these base learners are combined to give the ensemble predictor  $f(x)$ . In classification,  $f(x)$  is the most frequently predicted class:

$$f(x) = \arg \max_{y \in \mathcal{Y}} \sum_{j=1}^J I(y = h_j(x)) \tag{6}$$

In Random forests [15, 16] the  $j$ th base learner is a tree denoted  $h_j(X, \Theta_j)$ , where  $\Theta_j$  is a collection of random variables and the  $\Theta_j$ 's are independent for  $j = 1, \dots, J$ .

#### 4.4 Evaluation

Performance is measured using the confusion matrix as follows in Table 1.

**Table 1.** Confusion matrix.

	Predicted positive	Predicted negative
Positive actual	TP	FP
Negative actual	FN	TN

Based on the confusion matrix, evaluate the sentiment classifier through the measures are Accuracy, Precision, Recall, and F-Measure [17].

Accuracy is the proportion of correct predictions for both true positives and true negatives among the total number opinion quintuples of cases examined.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Precision is calculated between the number of true positives and the total number of true positives and false positives. If the result of this calculation has a value of 1, it represents the fact that all positively classified samples were true.

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Recall is the percentage of correct items selected. If the recall result is 1 it means all positive examples were found.

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

where  $TP$  is count of opinion quintuples correctly classified “Positive” sentiments,  $TN$  is count of opinion quintuples correctly classified “Negative” sentiments,  $FP$  is count of opinion quintuples incorrectly classified “Positive” sentiments,  $FN$  is count of opinion quintuples incorrectly classified “Negative” sentiments.

F-measure is used to evaluate the system’s overall performance by harmoniously combining the two metrics of recall and precision. The following formula defines F1 score.

$$F1 = \frac{2 * (Recall * Precision)}{Recall + Precision} \quad (10)$$

## 5 Opinion Summarization

Opinion summarization [2] based on the product aspect is made as follows:

*GENERAL.*

*Positive.* Total number of opinion holders who gave a positive opinion about the entity  $e$ .

*Negative.* Total number of opinion holders who gave a negative opinion about the entity  $e$ .

*Aspect  $a_1$ .*

*Positive.* Total number of opinion holders who gave a positive opinion about the aspect  $a_1$  of the entity  $e$ .

*Negative.* Total number of opinion holders who gave a negative opinion about the aspect  $a_1$  of the entity  $e$ .

*Aspect  $a_2$ .*

*Positive.* Total number of opinion holders who gave a positive opinion about the aspect  $a_2$  of the entity  $e$ .

*Negative.* Total number of opinion holders who gave a negative opinion about the aspect  $a_2$  of the entity  $e$ .

*Aspect  $a_n$ .*

*Positive.* Total number of opinion holders who gave a positive opinion about the aspect  $a_n$  of the entity  $e$ .

*Negative.* Total number of opinion holders who gave a negative opinion about the aspect  $a_n$  of the entity  $e$ .

where *GENERAL* deputizes the entity  $e$  itself and  $n$  is the total number of aspects of the entity  $e$ .

Given an opinion document  $D$ . The opinion summarization is modeled as follows in Fig. 4, where  $D$  is opinion document;  $e_i$  is entity;  $a_i$  is aspect of entity  $e_i$ ;  $o_i$  is opinion of aspect  $a_i$  of entity  $e_i$ .

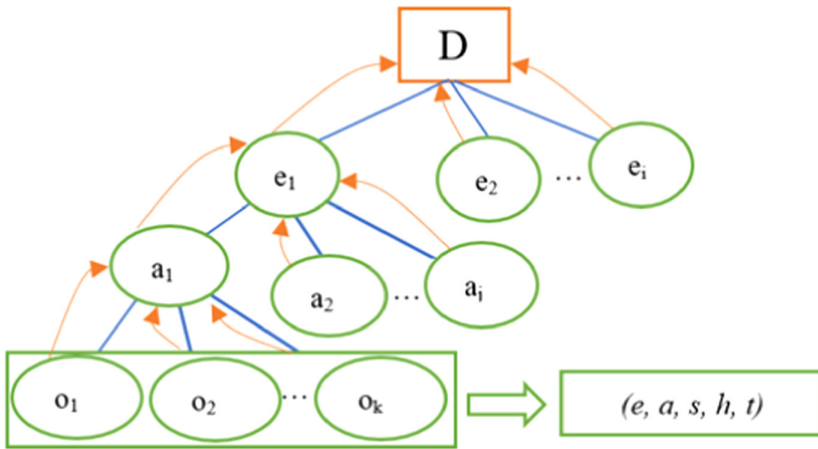


Fig. 4. Opinion hierarchy.

In the example of opinion summarization below. A structured opinion summary of aspects of a “Popchips” product such as “Flavor quality”, “Weight”, “Size” is described. There are 208 customer reviews expressing the positive opinion of “Flavor quality”, and only 4 negative. Meanwhile, the “Weight” aspect of the Popchips product has 128 positive reviews and only two negative reviews. Similar in terms of “Size” there are 228 positive reviews and 2 negative reviews. There are more positive reviews than negative reviews. This is common on food reviews. The *< individual review sentence >* link points to specific sentences or the entire review that gives a positive or negative comment about the aspect.

For an example of opinion summarization the “Popchips” product:

**Popchips.**

Aspect: **Flavor quality.**

Positive: 208

*< individual review sentence >*.

Negative: 4.  
 < *individual review sentence* >.

Aspect: **Weight**.  
 Positive: 128.  
 < *individual review sentence* >.

Negative: 2.  
 < *individual review sentence* >.

Aspect: **Size**.  
 Positive: 228.  
 < *individual review sentence* >.

Negative: 6.  
 < *individual review sentence* >.

## 6 Experiment

### 6.1 Data Used

The Amazon Food Reviews dataset [18] containing 568454 reviews is used to experimented. Each review contains the product identity, user identity, user name, user ratio (who found the reviews is helpful), product rating, review time, review summary, review text. Dataset details are described in Table 2.

**Table 2.** Amazon food reviews dataset.

Data set 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

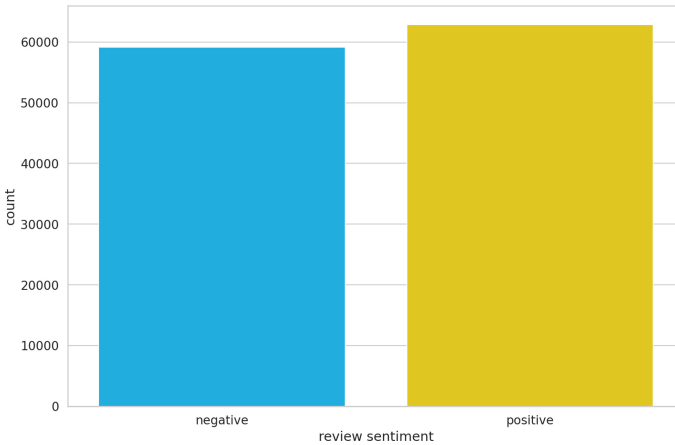
### 6.2 Preprocessing

The first one we prepossessing and remove Null values in the data set. The remaining data is 568411 reviews. The percentage according to the rating of the reviews in Table 3.

The second, processing balanced distribution of dataset, we split the dataset into positive and negative ratings with scores of 1 and 2 are negative, and scores of 3 and 4 are positive, and the score remove is 5. Figure 6. Illustration of the distribution of the number of positive and negative reviews. There are 59167 reviews that are negative, and 62856 reviews are positive out of a total of 568411 reviews.

**Table 3.** Detail of percentage according to the rating of the reviews.

Score	Total reviews	Percent
5	363111	63.88
4	80655	14.19
3	42638	7.5
2	29743	5.23
1	52264	9.19
Total	568411	100

**Fig. 5.** The distribution of the number of positive and negative reviews.

The next is randomly take 5411 reviews from the dataset 568411 reviews to perform the opinion discovery tasks in Sect. 3. The results obtained from the set of opinion quintuples, and then randomly divide the set of opinion quintuples into training opinion quintuplets and test opinion quintuplets in the ratio 8:2 respectively such in Table 4. It is used as input for opinion ensemble learning.

**Table 4.** Data train and test.

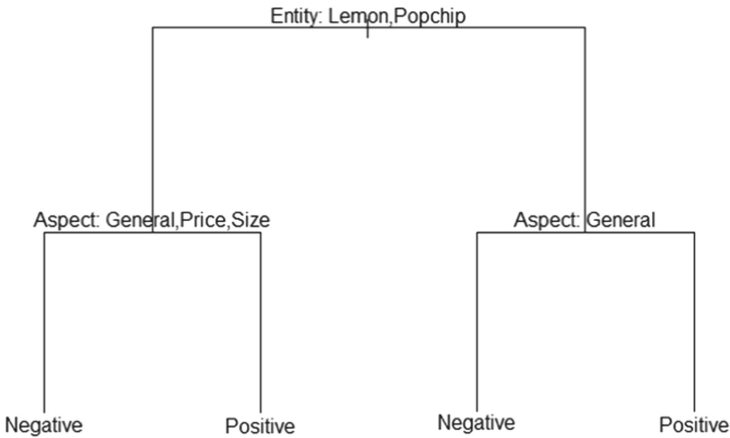
Opinion quintuples	Train	Test
8020	6416	1604

### 6.3 Tool Used

We proposed a technique implemented using R language [19]. Package “tree”<sup>1</sup>, and “randomForest”<sup>2</sup> were implemented for this experiment. In addition, we also proposed to implement package “NLP”<sup>3</sup> to perform text preprocessing data.

### 6.4 Scenario 1. Food Product Aspects

From the data set of food reviews, we have constructed a Random classification forest (or random forest) for opinion mining on food product aspects that have been commented on by customers. We first split the data into n bootstrap samples, then classify the weak learners as decision trees. We aggregate the classification by acquiring the value that occurs most often for each classification. The results of Random classification forest based on aspects of food products in Fig. 6.



**Fig. 6.** Random classification forest summary for product aspect.

**Table 5.** Detail of the measure of Random forest algorithm for product aspect.

Algorithm	Accuracy	Precision	Recall	F1
Random classification forest	0.805	0.8	0.76	0.78

The results on the measure of Random classification forest on the food reviews dataset with the participation of food product aspect have an accuracy of 0.805%, Precision is 0.8%, Recall is 0.76%, F1 is 0.78% in Table 5.

<sup>1</sup> <https://cran.r-project.org/web/packages/tree/>.  
<sup>2</sup> <https://cran.r-project.org/web/packages/randomForest/>.  
<sup>3</sup> <https://cran.r-project.org/web/packages/NLP/index.html>.

### 6.5 Scenario 2. Food Product Holders or Time Reviewers

In this scenario, we are interested in the important attributes of holder or time reviewer to consider whether sentiment orientation is positive or negative. The person evaluates his or her opinion and time to determine the direction of the point of view more accurately. Frequent reviewers with reviews are often more accurate than those who do not. The time reviewer is also important.

A review today will have a different opinion than previous reviews. Therefore, the scenario we propose is to have more important attributes of opinion evaluator and opinion evaluation time. The results of Random classification forest for food product holders or time reviewers are summarized in Fig. 7 and Fig. 8.

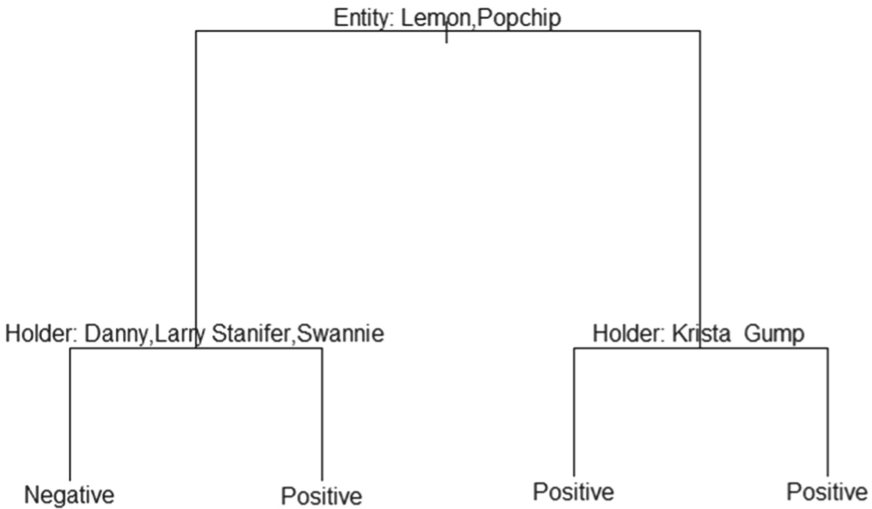
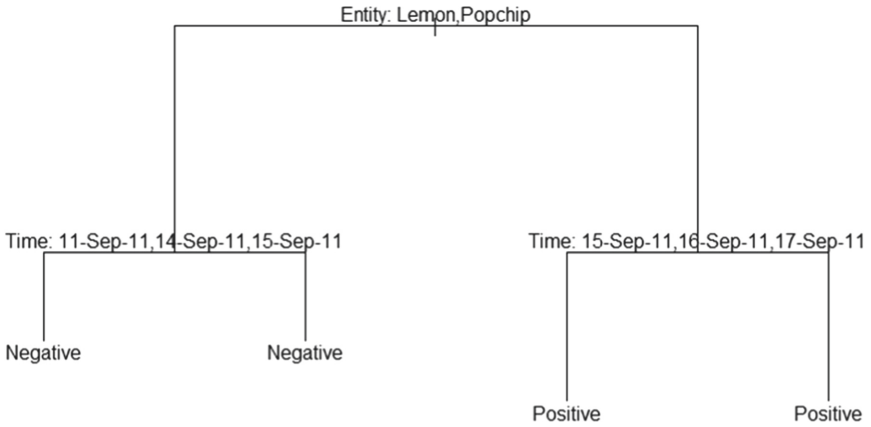


Fig. 7. Random classification forest summary for food product holders.

The results on the measure of Random classification forest on the food reviews dataset with the participation of food product holders or time reviewers have an accuracy of 0.82%, Precision is 0.85%, Recall is 0.67%, F1 is 0.75% in Table 6.



**Fig. 8.** Random classification forest summary for food product time reviewers.

**Table 6.** The measure of Random classification forest algorithm for food product holders or time reviewers.

Algorithm	Accuracy	Precision	Recall	F1
Random classification forest	0.82	0.85	0.67	0.75

## 7 Conclusion

We have approached ensemble learning modeling for opinion mining on food reviews. This model is based on a set of decision trees to classify affective orientation as positive, negative, or neutral. Each tree is built on a subset of random variables that are the attributes of the product name, product aspect, product reviewer, reviewer time, and the decision attribute which are sentiment attributes. Our experimental results on food reviews dataset based on two scenarios to explore important attributes in the construction of random forest classification have shown effectiveness.

## References

1. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012). <https://doi.org/10.2200/s00416ed1v01y201204hlt016>
2. Liu, B.: Sentiment Analysis: Mining Sentiments, Opinions, and Emotions, 2nd edn. Cambridge University Press, Cambridge (2020)
3. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pp. 79–86 (2002)
4. Abinash Tripathy, A., KumarRath, S.: Classification of sentimental reviews using machine learning techniques. Procedia Comput. Sci. **57**, 821–829 (2015). 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

5. Moret, B.M.E.: Decision trees and diagrams. *ACM Comput. Surv.* **14**(4), 593–623 (1982)
6. Zhou, Z.H.: *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, Boca Raton (2012)
7. Srivastava, R., Bhatia, M.: Ensemble methods for sentiment analysis of on-line micro-texts. In: *2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pp. 1–6 (2016)
8. Karthika, P., Murugeswari, R., Manoranjithem, R.: Sentiment analysis of social media network using random forest algorithm. In: *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, pp. 1–5 (2019)
9. Ahuja, M., Sangal, A.L.: Opinion mining and classification of music lyrics using supervised learning algorithms. In: *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, pp. 223–227 (2018)
10. Arnav Munshi, M.A., Thirunavukkarasu, K.: *Random Forest Application of Twitter Data Sentiment Analysis in Online Social Network Prediction*. Wiley Online Library (2021)
11. Ahmed, H., Awan, M., Khan, N., Yasin, A., Shehzad, F.: Sentiment analysis of online food reviews using big data analytics. *İlköğretim Online* **20**, 827–836 (2021)
12. Islam, N., Akter, N., Sattar, A.: Sentiment analysis on food review using machine learning approach. In: *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp. 157–164 (2021)
13. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. *Neural Comput.* **9**, 1545–1588 (1997)
14. LI, B., Friedman, J., Olshen, R., Stone, C.: *Classification and Regression Trees (CART)*, vol. 40 (1984)
15. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
16. Cutler, A., Cutler, D., Stevens, J.: Random forests. *Mach. Learn.* **45**, 157–176 (2011)
17. Sammut, C., Webb, G.I.: *Encyclopedia of Machine Learning*, 1st edn. Springer, Heidelberg (2011). <https://doi.org/10.1007/978-0-387-30164-8>
18. McAuley, J.J., Leskovec, J.: From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. *Association for Computing Machinery* (2013). <https://doi.org/10.1145/2488388.2488466>
19. Team, R.C.: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (2019)