Opinion Mining with Manifold Forests

Phuc Quang Tran[1], Hanh My Thi Le[2], and Hiep Xuan Huynh[3](✉)

[1] People's Police University, Ho Chi Minh City, Vietnam
tqphucth@gmail.com
[2] Faculty of Information Technology, The University of Danang - University of Science and Technology, Danang, Vietnam
ltmhanh@dut.udn.vn
[3] College of Information and Communication Technology, Can Tho University, Cantho, Vietnam
hxhiep@ctu.edu.vn

Abstract. Online reviews are becoming increasingly popular every day. They represent opinions and a wealth of information that can benefit organizations and individual consumers. However, studies on opinion mining have not focused much on classifying views according to the manifold to solve the problem of affinity between clusters of opinion, improving the accuracy, effectiveness, and generalizability of the modeling. In this paper, we have built an opinion mining framework with manifold forests to solve the influence of clusters of opinions based on the affinity between pairwise opinion points in each cluster and the relationship between different opinion clusters in large-scale data. In particular, we have focused on building a clustering trees ensemble and determining the affinity and distance of point pairs in feature space. Finally, the random forests are aggregated by ensemble methods such as stacking with a random forests classifier to identify opinion classification in reviews as either negative or positive. We used two datasets in the experiment to evaluate restaurants and hotels in two different scenarios, proving the effectiveness of the proposed model.

Keyword: Manifold forests · Clustering · Ensemble methods · Opinion mining

1 Introduction

With the explosive growth of global information [6], online applications on the internet are increasingly popular and effective for individuals and organizations. Increasing internet speeds, the appeal of social media sites, and e-commerce have resulted in a huge amount of informational data being reviewed online in the form of text. These reviews represent opinions [9] and a wealth of information that can benefit organizations and individual consumers. Among the techniques used in opinion mining, such as machine learning algorithms, lexicon-based approaches, and others, machine learning is commonly used in opinion mining. In machine learning algorithms, ensemble methods have many benefits to enhance the efficiency, performance, and generalizability of

P. Cong Vinh and N. Thanh Tung (Eds.): ICCASA 2023, LNICST 579, pp. 3–18, 2024.
https://doi.org/10.1007/978-3-031-58878-5_1

the model by dividing the original data into sub-datasets to match the basic learning models according to a certain method. The paper [10] presents an effective method for extracting opinion words to be used for opinion classification by ensemble learning [14]. Clustering algorithms are used to address the issue of processing unlabeled data. This paper [8] approaches the clustering method of product features for opinion mining. The proposed semi-supervised learning task uses the EM algorithm to solve the problem by improving labeled samples and allowing them to switch classes. The proposed method is evaluated effectively.

However, in the difficult case where opinion data comes from a combination of several non-linear dimensional manifolds, it is difficult to use conventional clustering algorithms for opinion clustering. Therefore, the use of manifold clustering [4] is a necessary solution to solve the problem of unsupervised machine learning in non-linear manifold space. In the article [3], the manifold clustering approach in deploying the manifold forest model has demonstrated that the manifold forest algorithm can close the gap with neural networks. In particular, the paper [1] has solved the general problem for clustering purposes by manifold clustering, which has superior results compared to worm subspace clustering based on autoencoders. The proposed Neural Manifold Clustering and Embedding (NMCE) method is closely related to and further understood with some self-supervised learning (SSL) methods. Currently, opinion mining studies have not focused on solving the problem of majority influence on the affinity between pairs of opinions in an opinion cluster or the affinity between different opinion clusters in large-scale opinion data.

In this paper, we propose an opinion mining framework with manifold forests to solve the problem of majority influence on the affinity between opinion points pairwise in an opinion cluster and the relationship between different opinion clusters in feature space. First, we focus on building a clustering tree ensemble. Second, we determine the affinity and distance of opinion point pairs in feature vector space. Third, we build independent random forests on the clusters. Finally, the random forests are synthesized by ensemble methods such as stacking with random forest classifiers to identify opinion classification in reviews as either negative or positive. We use two datasets of hotel and restaurant reviews to experiment with the model. The model's performance results are more accurate than the baselines, demonstrating the effectiveness of the proposed model. The framework of the proposed model is described in Fig. 1.

In addition to the content of Sect. 1, the paper is organized as follows: Sect. 2 discusses opinion modeling as the basis for opinion discovery in Sect. 3; Sect. 4 converts the results of opinion discovery into the opinion quintuples matrix; Sect. 5 presents the main content of opinion manifold forests; Sect. 6 evaluates the results of the model; Sect. 7 discusses opinion summarization; Sect. 8 experiments with the model on two different scenarios and discusses the results; Sect. 9 concludes the paper.

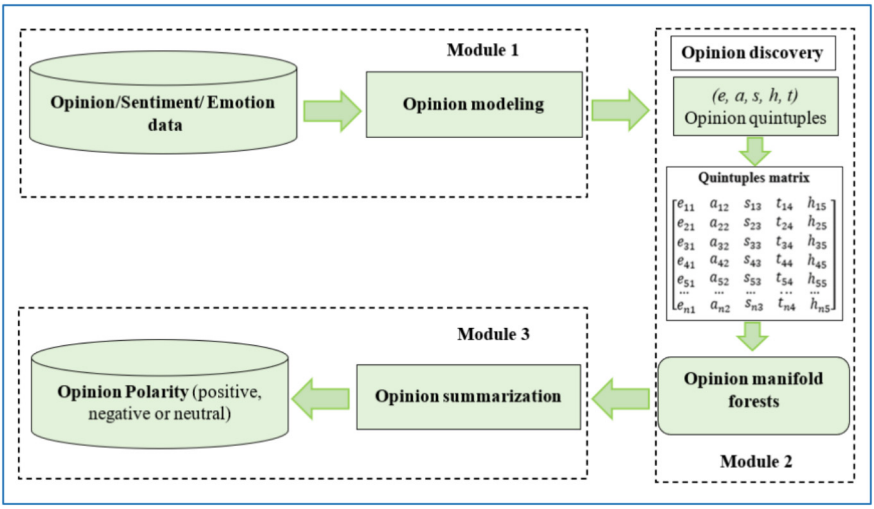

Fig. 1. The framework describing the proposed model

2 Opinion Modeling

2.1 Opinion

An opinion [9] modeled as a group of four components is called a quadruple (g, s, h, t), g is the opinion target, s is called the sentiment of the opinion towards the target g, and h is the person or organization giving opinion is called an opinion holder, t is the time to give the opinion of the person or organization. An opinion holder is an individual or organization that states an opinion. The time of opinion is the posting time or statement of the opinion holder.

Opinion Target. [9] The target of opinion or sentiment is the sentiment expressed to an entity or an aspect of the entity.

Sentiment of Opinion. [9] Sentiment of opinion is modeled as a triple quadruple (y, o, i). Sentiment types y can be classified into several categories based on linguistics, psychology, and classification based on consumer research. On the basis of consumer research is divided into two categories: rational emotions and emotional emotions. The sentiment of opinion is expressed through sentiment orientation o or opinion polarity can be negative, positive, or neutral. The sentiment of the opinion also depends on the intensity of the sentiment i, i.e. the different strengths and weaknesses of the sentiment. In practical application, it is possible to represent sentiment intensity as a number of discrete ratings based on two sentiment types such as a rating of 1 to 5 stars.

2.2 Simplify Opinion Definition

An opinion [9] can be defined simply and in more detail as a quintuple (e, a, s, h, t) where e is the entity to which the target is directed or the target is directed to the aspect a of

the entity e, s is a sentimental orientation or polarity of opinion whose value is negative, positive or neutral, possibly also a 1 to 5-star rating. If sentiment reflects directly to the entity, the aspect represented is GENERAL. In this case e and a also express sentiment target.

In addition to the five components quintuple (e, a, s, h, t) of an opinion, there are reasons and qualifiers. A reason for an opinion [9] is the cause or explanation of the opinion. A qualifier [9] of an opinion limits or modifies the meaning of the opinion.

2.3 Opinion Entity

An opinion entity [9] e is represented by itself as a whole and a finite set of opinion aspects $A = \{a_1, a_2, \ldots, a_s\}$. Each aspect $a \in A$ of entity e can be expressed with any one of a finite set of its aspect expressions $\{ae_1, ae_2, \ldots, ae_n\}$.

2.4 Opinion Document

An opinion document [9] D includes a finite set of opinion entities $\{e_1, e_2, \ldots, e_r\}$ and a subset of opinion aspects $\{a_1, a_2, \ldots, a_s\}$ of each opinion entity. The opinion is derived from a finite set of the opinion holder $\{h_1, h_2, \ldots, h_p\}$ and at a particular time t.

3 Opinion Discovery

Given an opinion document D, opinion discovery on opinion document D is performed on the order of eight tasks as follows [9]:

First, extract the expressions of entities in document D and group similar entities and group entities into clusters or classifications. Each entity represents an entity clustering expression. Second, extract the aspect expressions in document D and group the aspects into clusters similar to the first task. Third, extract the holder's expression of each opinion from reviews or structured data and group them. Fourth, extract the posting times of each opinion and normalize the times by different formats. Fifth, classifies the perspective of the aspect by identifying the aspect or entity that has a positive, negative, or neutral opinion. Sixth, synthesize the above tasks to form all groups of opinions quintuble (e, a, s, h, t). Seventh, to extract the opinion reasons for each opinion and group the synonymous reason into a cluster. Each reason expression represents a group of opinion reasons. The final, extract the opinion qualifier expression for each opinion and do the same as the seventh task.

4 Opinion Quintubles Matrix

Given an opinion quintuple (e, a, s, h, t), where e, a, s, h, t are the features of the opinion whose values are respectively $(e_{n1}, a_{n2}, s_{n3}, h_{n4}, t_{n5})$, where n number of sample values. The value of the opinion feature can be a discrete value, a continuous value, or a mixed value. Convert opinion feature values to numeric values to get an matrix O : n x 5, with n number of sample values of opinion features represented as follows (Fig. 2).

The matrix $\mathcal{O}$ is the input data to train the opinion manifold forest model that we present below.

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

Fig. 2. Matrix of Quintubles

5 Opinion Manifold Forests

An opinion manifold forest is an ensemble collection of clustering trees that simultaneously considers the relationship between points in the opinion feature space. Therefore, the opinion manifold forests model needs to estimate the affinity or distance between opinion points to be able to preserve the distance of those data points after mapping. In Fig. 3 proposed the manifold forests for opinion polarity/opinion classification.

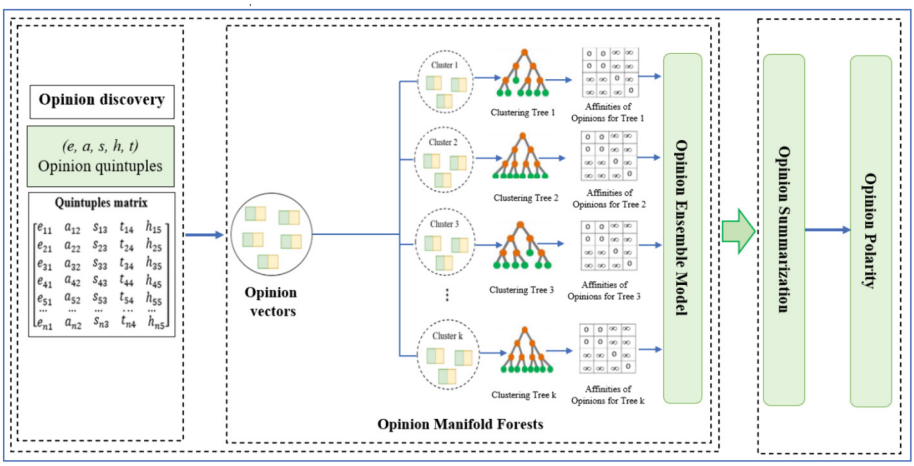

Fig. 3. The proposed model of the manifold forests for opinion polarity

5.1 Opinion Vectors

In the opinion feature space for opinion feature vectors $O = (e, a, s, h, t) \in \mathbb{R}^5$, each vector $o \in O$ is responsible for representing a point in the feature space and each vector opinion has five dimensions representing features for an opinion, each feature has a numeric data type and is one dimensional. These opinion feature vectors are used to train the opinion manifold forest model.

5.2 Opinion Manifold Forests Model

Given a set of k opinion samples $O = \{o_1, o_2, \ldots, o_k\}$ unlabled with $o_i \in \{R\}^d$, find a mapping $f : \{R\}^d \rightarrow \{R\}^{d'}, f(o_i) = o'i$ so that $d' \ll d$ and preserve affinity or distance related points in feature vector space. Each opinion input sample $o_i = (e, a, s, h, t) \in \{R\}^5$. The output is sentiment s $\in$ {positive, negative}

The opinion manifold forest model is trained using the maximization of random nodes and trains the forest by maximizing the information gain measure. So, the process of splitting the jth node for optimization is done by the objective function as follows

$$\theta_j = \underset{\theta \in \tau_j}{\operatorname{argmax}} I\left(O_j, \theta\right) \tag{1}$$

with I is the continuous information gain as in the opinion density forest model [16]

$$I\left(O_j, \theta\right) = \log\left(\left|\Lambda\left(O_j\right)\right|\right) - \sum_{i \in \{L,R\}} \frac{\left|O_j^i\right|}{\left|O_j\right|} \log\left(\left|\Lambda\left(O_j^i\right)\right|\right) \tag{2}$$

where O_j is the set of opinion samples that split the jth node, O_j^i is the set of opinion samples on the path to the left and right of the jth node, $\Lambda\left(O_j^i\right)$ is the covariance matrix at node split jth.

Different from opinion density forests [16], the opinion manifold forest model estimates a measure of the affinity between opinion data points to maintain the distance between those data points after performing mapping using random forests to define opinion affinity. At the leaves of a clustering tree t defines a partition of the input opinion points.

$$l(o) : \{R\}^5 \rightarrow L \subset \{N\} \tag{3}$$

with l as the leaf node index and L as the set of all leaves in a tree. Each clustering tree t can compute the affinity matrix $\mathbb{W}^t : \mathtt{k} \times \mathtt{k}$ with

$$W_{ij}^t = e^{-Q^t\left(o_i, o_j\right)} \tag{4}$$

where $\mathtt{Q}$ is the distance determined by the binary affinity of the pair $\left(o_i, o_j\right)$ of opinions.

$$Q^t\left(o_i, o_j\right) = \begin{cases} 0 & if \;\; l(o_i) = l\left(o_j\right) \\ \infty & otherwise \end{cases} \tag{5}$$

where $d_{ij} = o_i - o_j$, and $\Lambda_{l(o_i)}$ is the covariance matrix linked to the leaf node by the point o_i.

Given a tree t and two o_i and o_j, if points o_i and o_j end up in the same cluster (leaf), then assign affinity equal to 1, distance equal to zero for the pair $\left(o_i, o_j\right)$. Otherwise, assign affinity equal to zero, distance equal to ∞.

The opinion affinity matrix of the manifold forests is the ensemble affinity matrix of the trees of the entire forest of size T and is calculated by averaging [12] affinity matrix $\mathbb{W}^t$ from single opinion clustering trees

$$W = \frac{1}{T}\sum_{t=1}^{T} W^t \tag{6}$$

6 Evaluation

The proposed model is evaluated through Table 1 called the opinion confusion matrix, which consists of four different combinations of opinion predictor value and actual opinion value. The classification performance of the model in this case is evaluated as either a negative or positive classifier [15]:

Table 1. Opinion confusion matrix

	Opinion predictor positive	Opinion predictor negative
Actual opinion positive	True Positive (TP)	False Positive (FP)
Actual opinion negative	False Negative (FN)	True Negative (TN)

Where True Positive (TP) represents the number of opinion samples with true positive opinion values that are predicted to be true positive. False Positives (FP) represent the number of opinion samples with true positive opinion values predicted that are not true positive. False Negative (FN) represents the number of opinion samples whose true negative opinion values are predicted to be neither true negative. True Negative (TN) represents the number of opinion samples with true negative opinion values that are predicted to be true negative.

Measures of accuracy, precision, recall, and F1 are used to evaluate the opinion classification performance of the proposed model.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

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

7 Opinion Summarization

Opinion summarization is an aggregation of user opinions expressed in reviews online. The major opinion summarization work often takes the entity-centric aspect of the entity as a hierarchy to create a specific summary consisting of documents related to an entity or an aspect of the entity. Such opinion summaries are called aspect-based opinion summaries or feature-based opinion summaries [9]. We aggregate opinions on features using summaries of different products so that consumer opinions can be compared to competing products. Figure 4 shows a visualization of customer opinions of different entities.

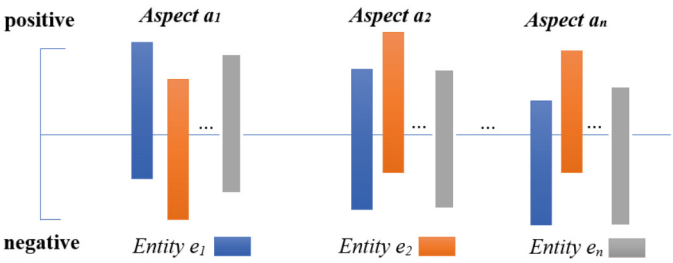

Fig. 4. Visualization of opinions on different entities

Where, $e = \{e_1, e_2, \ldots, e_n\}$ is the set of entities that are the names of the products corresponding to the set $a = \{a_1, a_2, \ldots, a_n\}$ aspects of the entity e. Each entity can also have many different aspects. The entity opinion or aspect of the entity is positive or negative.

In each of the different colored bars, Fig. 4 shows the percentage of reviews that express a negative or positive opinion on a horizontal axis about a certain aspect. With the above opinion summarization, it is very convenient for customers to visually observe to know the strengths and weaknesses of each product to make product selection decisions.

8 Experiment

8.1 Data Used

In this experiment, we used two datasets YelpHotelData and YelpResData [6] as the basis for training the proposed model.

The YelpHotelData dataset is a database containing content related to hotel opinions, hotel reviewers, and hotel reviews. This dataset has been extracted, and processed into structured data containing important information from reviews related to aspects of the hotel such as credit card payment services *(AcceptsCreditCards)*, in relation to the price service *(PriceRange)*, and the hotel's Wi-Fi facility *(WiFi)*, opinion holder and time of review are also listed in the data. The details of the YelpHotelData dataset are described in detail in Tables 2 and 3.

The YelpResData dataset contains content related to opinion restaurants, restaurant reviewers, and restaurant reviews. Similar to the YelpHotelData dataset, this dataset contains more restaurant aspects than the YelpHotelData dataset. The details of the YelpHotelData dataset are described in detail in Tables 4 and 5.

Table 2. The details of tables of the YelpHotelData dataset

Tables	Total columns	Total rows
Hotel	13	283086
Reviewer	13	5123
Review	10	688329

Table 3. The details of columns of the YelpHotelData dataset

Hotel	Reviewer	Review
HotelID	ReviewID	Date
Name	Name	ReviewID
Location	Location	ReviewerID
ReviewCount	YelpJoinDate	ReviewContent
Rating	FriendCount	Rating
Categories	ReviewCount	UsefulCount
Address	FirstCount	CoolCount
AcceptsCreditCards	UsefulCount	FunnyCount
PriceRange	CoolCount	Flagged
WiFi	FunnyCount	HotelID
Website	ComplimentCount	
PhoneNumber	TipCount	
FilReviewCount	fanCount	

Table 4. The details of tables of the YelpResData dataset

Tables	Total columns	Total rows
Restaurant	30	242652
Reviewer	13	16941
Review	10	788471

Table 5. The details of columns of the YelpResData dataset

Restaurant	Restaurant	Reviewer	Review
RestaurantID	Delivery	ReviewID	Date
Name	Takeout	Name	ReviewID
Location	WaiterService	Location	ReviewerID
ReviewCount	OutdoorSeating	YelpJoinDate	ReviewContent
Rating	WiFi	FriendCount	Rating
Categories	GoodFor	ReviewCount	UsefulCount
Address	Alcohol	FirstCount	CoolCount
Hours	NoiselLevel	UsefulCount	FunnyCount
GoodforKids	Ambience	CoolCount	Flagged
AcceptsCreditCards	HasTV	FunnyCount	RestaurantID
Parking	Caters	ComplimentCount	
Attire	WheelchairAccessible	TipCount	
GoodforGroups	WebSite	FanCount	
PriceRange	PhoneNumber		
TakesReservations	FilReviewCount		

8.2 Preprocessing

The YelpHotelData dataset is preprocessed from the original data converting the data into a five-component structure of the opinion as *(Hotel names, Hotel aspects, Opinion ratings, Hotel reviewers, Reviewer times)* corresponding to the opinion quintuple (e, a, s, h, t) have been suggested above. The value of the five components e, a, s, h, t is converted to a numeric value. The details of data conversion results have 882,474 opinion quintuples are reported in Table 6.

Table 6. The details of hotel data conversion

The features of hotel review	Different values of features
Hotel names	from 1 to 123461
Hotel aspects	from 1 to 3
Hotel reviewers	from 1 to 4596
Reviewer times	from 1 to 4382
Opinion ratings	from 1 to 5

Similarly, the YelpResData dataset is preprocessed into a five-component structure of the opinion as *(Restaurant names, Restaurant aspects, Opinion ratings, Restaurant*

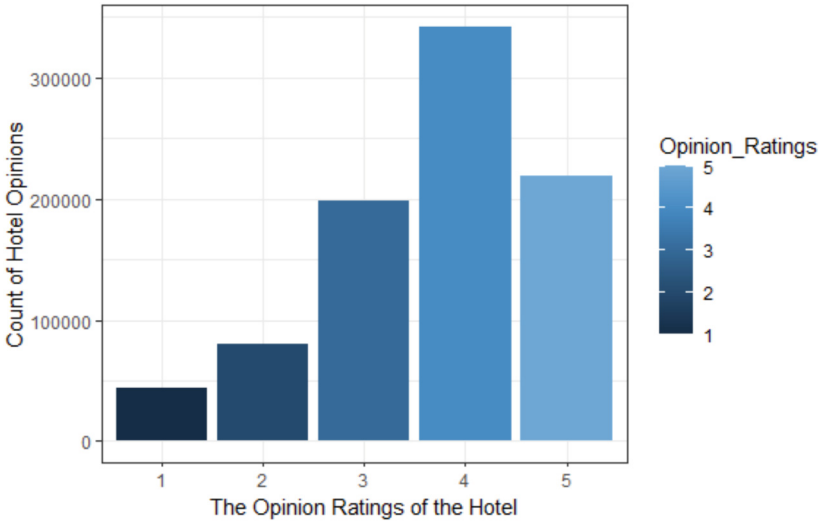

Fig. 5. Distribution of opinion ratings about the hotel

reviewers, Restaurant times). The distribution of opinion ratings about the restaurant in Fig. 6. We explored nineteen aspects that influence the quality of the restaurant such as GoodforKids, AcceptsCreditCards, Parking, Attire, GoodforGroups, PriceRange, TakesReservations, Delivery, Takeout, WaiterService, OutdoorSeating, WiFi, GoodFor, Alcohol, NoiselLevel, HasTV, Ambience, Caters, and WheelchairAccessible. The results of the initial data conversion have 1,479,1500 sets of opinion quintuples (Fig. 5).

Table 7. The details of restaurant data conversion

The features of restaurant	Different values of features
Restaurant names	from 1 to 184167
Restaurant aspects	from 1 to 19
Restaurant reviewers	from 1 to 12943
Restaurant times	from 1 to 4541
Opinion ratings	from 1 to 5

The ratings of the opinions of the reviews from 1 to 5 stars correspond to the five levels of opinion polarization as “very negative”, “negative”, “neutral”, “positive”, and “very positive”. In the proposed model experiment, we propose to polarize opinions at two levels “negative” and “positive”. Therefore, from the five polarizing perspectives we convert to two opinions polarity by combining the “very negative” and “negative” polarizing into “negative”; “positive”, and “very positive” to “positive” and remove the polarization of the “neutral” opinion. The hotel opinion rating conversion results have 560,769 negative samples and 321,705 positive samples in Fig. 7. The restaurant

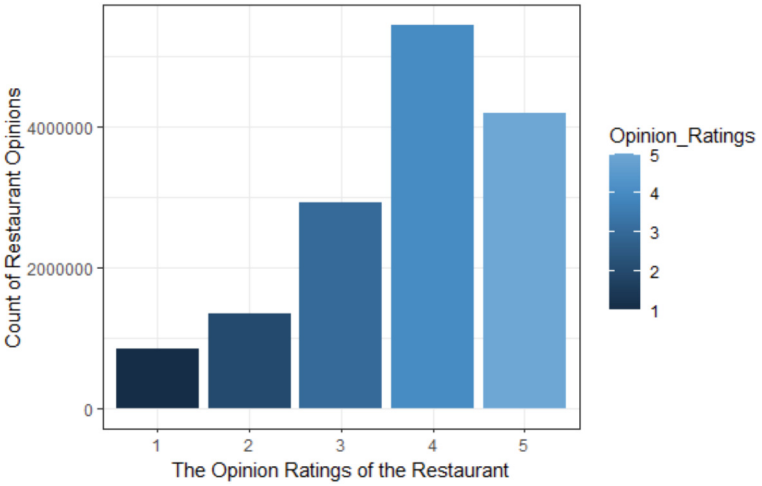

Fig. 6. Distribution of opinion ratings about the restaurant

opinion rating conversion results have 5,140,963 negative samples and 9,650,537 positive samples in Fig. 8 (Table 7).

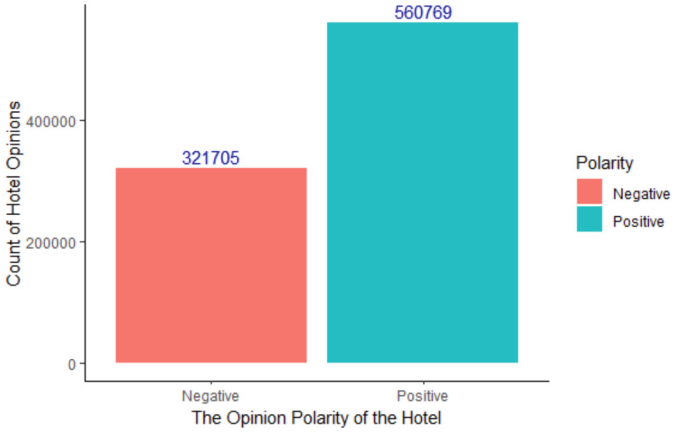

Fig. 7. The distribution of hotel opinions

8.3 Tool Used

In order to conduct experiments, we used R language to build the proposed model and integrated some main libraries into the model such as library(celltrackR) [10], library(stacks), library(randomForest), library(h2o), library(dplyr), and several support libraries for data processing and visualization.

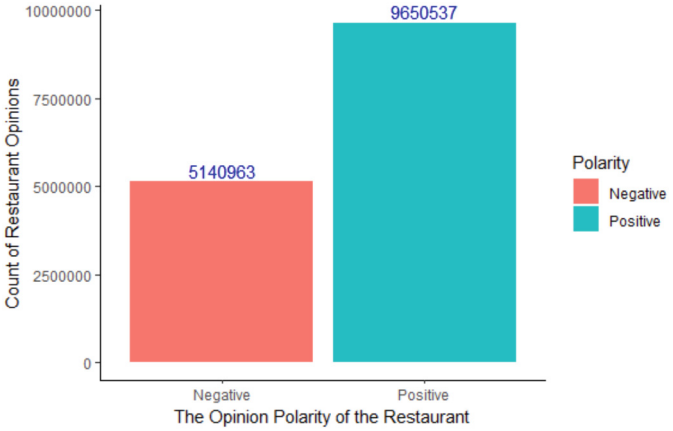

Fig. 8. The distribution of restaurant opinions

8.4 Scenario 1. Manifold Forests for Hotel Opinion.

The proposed model is trained on the YelpHotelData dataset that has been preprocessed to opinion quintubles and converted to feature vector space as input for training the model based on the following steps:

First, is to split the initial data set into a training dataset, validation dataset, and test dataset according to the following ratio tương ứng là 60%, 20%, 20%. The opinion training dataset is omitted from the labeling to perform clustering. As a result, three clusters are created.

Running three random forest models on 3 clusters such as the random forest 1 (RF1), the random forest 2 (RF2), and random forest 3 (RF3) with cross-validation method (k-folds = 5).

Get prediction results of 3 models RF1, RF2, RF3 from valid set (20%).

Build a random forest (RF) model to stack from the prediction results of three models such as RF1, RF2, RF3 to form a manifold forest (MF).

Comparing the accuracy of each model with the manifold forest (MF) on the test data set.

The performance in Table 8 of the manifold forest model on the hotel training set for accuracy, precision, recall, and F1 measures are 94%, 90%, 70%, and 79%, respectively. The testing set is 78%, 74%, 62%, and 79% for accuracy, precision, recall, and F1, respectively.

Table 8. Evaluate manifold forests for hotel opinion dataset

Hotel dataset	Accuracy	Precision	Recall	F1
Training set	0.94	0.90	0.70	0.79
Testing set	0.78	0.74	0.62	0.67

The comparison results are shown in Fig. 9. We can see the performance for each individual model. The accuracy of the random forest 3 (RF 3) for the training set is 90%,which is the lowest. The accuracy of RF 1, and RF 2 are 91%, and 92% respectively. The accuracy of MF is 94% for the training set, which is the highest.

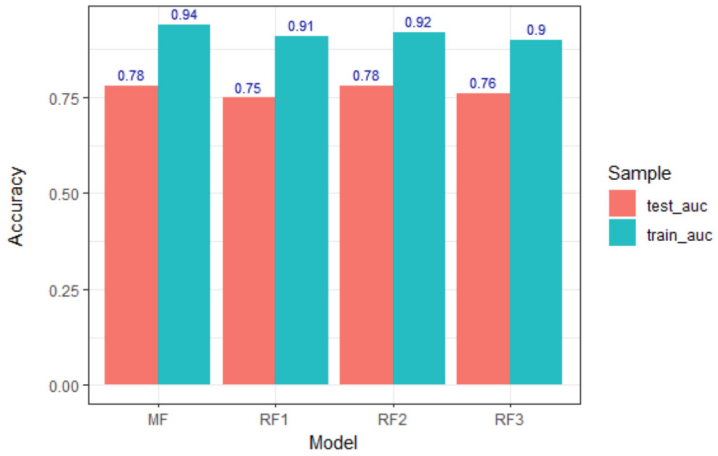

Fig. 9. The accuracy of the random forests 1, the random forests 2, the random forests 3, and manifold forests (MF) for the hotel opinion dataset.

8.5 Scenario 2. Manifold Forest for Restaurant Opinion.

Similar to scenario 1, we built the manifold forest for the restaurant opinion dataset by applying the following steps:

First, we have divided data set into a training dataset, validation dataset, and test dataset with ratios of 60%, 20%, 20% respectively. Then perform clustering on the unlabeled training dataset. As a result, four clusters are created. We build four independent random forests including RF1, RF2, RF3, and RF4 respectively on four clusters with the cross-validation method (k-folds = 10).

Get prediction results of four models RF1, RF2, RF3, and RF4 from the valid set (20%).

Applying random forest (RF) to stack from the prediction results of four models such as RF1, RF2, RF3, and RF4 to form a manifold forest (MF).

The manifold forests on the restaurant opinion training set in Table 9 have a performance of 95%, 92%, 71%, and 80% for accuracy, precision, recall, and F1measures, respectively. The testing set is 82%, 75%, 67%, and 71%, for accuracy, precision, recall, and F1 measures, respectively

The accuracy of RF 1, RF 2, RF 3 are 93%, 94%, 92% for the training set, respectively. In which, the accuracy of RF 3 is the lowest. The accuracy of MF is 95%, which is the highest for the training set in Fig. 10.

Table 9. Evaluate manifold forests for the restaurant opinion dataset

Restaurant dataset	Accuracy	Precision	Recall	F1
Training set	0.95	0.92	0.71	0.80
Testing set	0.82	0.75	0.67	0.71

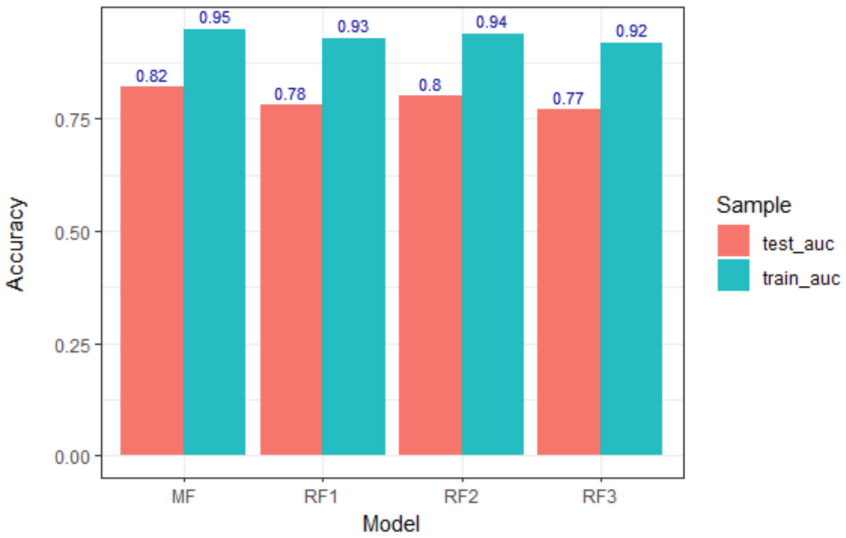

Fig. 10. The accuracy of forests for the restaurant opinion dataset

9 Conclusion

In this study, we propose to build an opinion mining model with the manifold forests approach to solve the problem of clustering opinions according to the affinity between different clusters of opinion in large-scale opinion data. In particular, we focus on building a random forest model on the clusters to determine the affinity of the opinion data and aggregate the forests by ensemble methods to identify the opinion classification in reviews as either positive or negative. We conducted experiments using a dataset of hotel and restaurant reviews. The results show that manifold forests can accurately estimate opinion classification, and the use of ensemble learning, such as the stacking algorithm, achieves the best results. This means that ensemble learning can improve the accuracy and efficiency of opinion classification in hotel and restaurant reviews. This finding could be useful for opinion mining to choose an effective model.

Acknowledgments. The authors gratefully acknowledge the support of distinguished professor Bing Liu, Department of Computer Science University of Illinois at Chicago (UIC) for sharing the dataset to conduct this study.

References

1. Li, Z., Chen, Y., LeCun, Y., Sommer, F.T.: Neural manifold clustering and embedding. In: Proceedings of the 10th International Conference on Learning Representations (ICLR) (2022)
2. Shiebler, D.: Functorial manifold learning and overlapping clustering. In: Proceedings of Machine Learning Research, vol 132, pp.1–20 (2021)
3. Perry, R., Tomita, T. M., Patsolic, J., Falk, B., Vogelstein, J.T.: Manifold forests: closing the gap on neural networks. In: Proceedings of the 8th International Conference on Learning Representations (ICLR) (2019)
4. Souvenir, R., Pless, R.: Manifold clustering. In: Proceedings of the 10th International Conference on Computer Vision (ICCV), pp. 648–653 (2005)
5. Ester, M., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD), ACM, 226–231 (2016)
6. Bordoloi, M., Biswas, S.K.: Sentiment analysis: a survey on design framework, applications and future scopes. Artif. Intell. Rev. (2023). https://doi.org/10.1007/s10462-023-10442-2
7. Mukherjee, A., Venkataraman, V., Liu B., Glance N.: What yelp fake review filter might be doing. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), Boston, USA (2013)
8. Zhai, Z., Liu, B., Xu, H., Jia, P.: Clustering product features for opinion mining. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM'11), pp. 347–354 (2011)
9. Liu, B.: Sentiment Analysis: Mining Sentiments, Opinions, and Emotions. 2nd edn. Cambridge University Press, Cambridge (2020)
10. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
11. Tran, P.Q., Trieu, N.T., Dao, N.V., Nguyen, H.T., Huynh, H.X.: Effective opinion words extraction for food reviews classification. Int. J. Adv. Comput. Sci. Appl. **11**(7), 421–426 (2020)
12. Gautam K.: Ensemble Methods for Machine Learning. Manning Publications Co., Shelter Island, NY (2023)
13. Wortel, I.M.N., Liu, A.Y., Dannenberg, K., Berry, J.C., Miller, M.J., Textor, J .: CelltrackR: an R package for fast and flexible analysis of immune cell migration data. ImmunoInform. 1–2 (2021). https://ingewortel.github.io/celltrackR/
14. Tran, P.Q., Nguyen, H.T., Le, H.M.T., Huynh, H.X.: Ensemble learning for mining opinions on food reviews. In proceedings of the International Conference on Context-Aware Systems and Applications(ICCASA 2021), pp 56–70 (2021)
15. Powers, D.M.W.: Evaluation: from precision, recall and f-measure to roc., informedness, markedness and correlation. J. Mach. Learn. Technol. **2**, 37–63 (2011)
16. Tran, P.Q., Ha, D.N.L., Le, H.T.M., Huynh, H.X: Opinion mining with density forests. J. EAI Endorsed Trans. Context-aware Syst. Appl. **9**(1), (2023)