Aspect based approach for document level sentiment classification of movie reviews

Abinash Tripathy[1], Ch. Chakradhar Rao[2], Panchanand Jha[3] and Gandi Satyanarayana[4]

[1,2]Department of Computer Science and Engineering, Raghu Engineering College, Visakhapatnam

[3]Department of Mechanical Engineering, Raghu Engineering College, Visakhapatnam

[4]Department of Computer Science and Engineering, Avanthi Institute of Engineering and Technology, Visakhapatnam

Abstract

In recent times, with the increase in the use of social media and online trading sites, the customers or the users prefer to share their views about the product they have brought. These comments are generalized for the product but in some cases like electronics products or mobile, the customers review the product based on some aspects like mobiles, camera quality, sound, the weight of mobile. In this paper, the aspect-based sentiment analysis approach is adopted to find out the sentiment of the reviews. The reviews under analysis are initially processed based on an aspect and then the classification is done using a machine learning technique. In this paper, Support Vector Machine (SVM) is considered for analysis as this method has given a better result in the field of sentiment analysis. Finally, the result is accessed by using different evaluation parameters to validate the quality of the work done in the paper.

Keywords: Aspect, Aspect based sentiment analysis, Support Vector Machine, Evaluation parameters

1. Introduction

The customers always feel happy to share information about a product they have bought and also share their anger or dislike about the product in the social media. Every on-line site requests their customer to provide their feedback about the product they have bought, which is publicly available and whoever wants to buy the product, they can go for it. The dislike of the customers is also handled by the sellers through proper mechanism. Thus, the credit of the product increases and it became popularize among the customers. For electronics products, such as for mobile, the choice of the product depends up on criteria like camera, display, storage capacity, size of the mobile etc. and the customer not only make their search based on that but also buy the products based on their choice criteria. So, Sentiment analysis can be broadly informed as a technique to collect the comments or feedback of the customers from a reliable source and then process those using different techniques and finally, provide an information that will be useful to both the customers and seller for future works [1].

The reviews that are considered for analysis comes in many different forms. Some reviews prefer to write their feedbacks in paragraphs, some confined to one or two sentence, some highlight one aspect of the product like `` movie is romantic" or ``movie is full of suspense", and in some case product comparison is carried out. Feldman has gone through thetypes of sentiment analysis and suggested the categories based on the granularity level [2]. He has suggested the following categories for the sentiment analysis.

- Document level: Some reviews are not confined to one or two sentences. These comments are in paragraphs. In order to process these reviews, the whole paragraphs are considered as a single unit and the polarity of the document is provided for the whole.
- Sentence level: The reviews provided by some reviewers are confined to a single statement or at max two to three statement. In such case, these reviews are analysed based on sentence level categories. The sentiment values of each sentence are calculated separately and finally combined together to obtain the sentiment of reviews.
- Comparative analysis: In some cases, instead of giving feedback on any product, customers prefer to compare that with another one, which is available in the market already and gain some popularity among

the users. In such cases, the customer who already has an experience of using the old product can think about the recent product to buy or not.

- Aspect based: Every product has some properties, for example, when a cloth is considered, its material, colour, quality; for mobiles the screen size, memory size, camera quality etc. In such conditions the customer buy their product based on the aspect only, they do not take in to consider the other criteria i.e., how good or how bad they are.

In the present study, the combination of both aspects based and document level sentiment analysis is considered. In this paper, movie review dataset is considered for analysis and the aspect considered here is ``Romantic".

The reviews can be processed using many different approaches such as machine learning, graphical approach [3], inductive based approach [3]. In this paper, machine learning approach is preferred as it is comparatively easier and faster approach for processing of the reviews. Machine learning (ML) is one the recent tool used for processing of reviews or comments. It can be of many different types. The commonly used ML techniques by different authors can be stated as follows:

- Supervised approach: It is one of the most preferred ML approaches for analysis. In this type of approach, the comments are labelled beforehand. Using ML approaches, polarity of the comments is predicted and matched with the original one. If both are matches then the proposed approach has a higher accuracy or the proposed approach is not efficient [4]. This approach is preferred to classify the comments into preassigned groups.
- Unsupervised approach: In this type of approach, there is no label associated with the reviews and thus, processing approach is different. Here, based on the features of the reviews, they are grouped together and the approach is called as clustering approach [5].
- Semi-supervised approach: It is observed that processing the labelled reviews in quite easier in compare to the unlabelled reviews. Thus, this approach makes an attempt to provide label to the unlabelled reviews collected from different sources for analysis[6]. This approach adopted in this technique is quite new as compared to supervised and unsupervised approach

The present study deals with the supervised learning approach.

The proposed chapter is organized as follows: The section 2 informs about the literature review on the topic of Semi-supervised analysis. Section 3 discusses information about adopted methodologies in the paper. Section 4 provides the information about approach adopted and result analysis. Finally, the concluding information and suggested future work of the chapter suggested in Section 5.

2. Literature Survey

Sastry and Karthick have proposed an unsupervised framework for item rating prediction [7]. This framework uses Latent Dirichlet Allocation (LDA) to extract topics from reviews and labels each topic with the aspect name. They use PageRank to estimate the aspect importance. The use of PageRank assigns proper weights to the aspects, when compared to LDA. Deep Neural Networks are used in Deep connection to extract latent aspects from reviews and predict ratings using these aspects. The overall rating of the product is calculated as the weighted summation of the latent aspect ratings and their importance scores. This approach produced an accuracy of 69% in aspect labelling and 62% in sentiment labelling and low RMSE error which is better when compared to the other unsupervised models.

Lin *et.al.* have suggested an aspect-based sentiment analysis with hybrid attention that uses attention approach instead of convolution or recurrent style of processing [8]. This approach considers a particular aspect of any product based on the sentiment of the user, these aspects may be any property, feature or working style of the product. For the purpose of analysis, they have considered three different datasets such as car, SemEval2014 Task 4 datasets and Twitter dataset. Their proposed SANET_BERT approach helps in increase of accuracy value by 9.33 % and F1 score by 19.17 % on average. They observed that SANET based approach is more efficient than the previous models proposed and the time complexity is found to be quite less. In order to increase the quality of their approach they collect information from Sent WordNet and the words mostly used in the field of service during the analysis of reviews and found out a better outcome.

Zhang *et.al.* have combined both machine and deep learning to prepare a model, which is ensemble based for sentiment classification [9]. Their proposed approach based on user reviews about restaurants based on location and user-based application for 5G networks. They faced a challenge to handle the reviews collected from users in 5G, they faced the issues like complex design of the model, unavailability of labelled reviews. Their proposed approached in granularity based. For single level granularity, they obtained an accuracy of around 96.78% and the F-measure value of 0.863. They proposed an aspect-based system, which is multi aspect based to obtained a reasonable number of labelled reviews. Their proposed approach has shown a better result in compare to that of the traditional approach used for sentiment classification.

Sindhu *et.al.* have proposed sentiment classification based on aspect to obtain the extremity of the target sense [10]. They have used deep learning approach for classification, which is now-a-days popularly used for sentiment analysis. For this purpose, they have combined the CNN and bidirectional gate recurrent unit (BiGRU). Their approach selects features from the test reviews using BiGRU and the static features are obtained using CNN. Finally, sigmoid classifier is used for classification. Their proposed approach has obtained an accuracy of 78.14 on the review dataset collected on restaurants. They have used different machine learning techniques for the analysis and found out that the proposed approach has performed a good job in the field of emotional classification work.

Li *et.al.* Proposed an approach of sentiment classification based on aspect to find out the spam groups on different topics [11]. They have considered a dataset collected from business-to-customer websites like JD.com and TMALL.com. Their dataset consists of 14821 reviews from JD and 29986 reviews from TMALL. They suggested GSDNT method for finding out the spam groups based on the nominated reviews collected from both the sites. Their proposed approach consisting of three distinguished phases like Pre-processing to define equivalent groups, mining recommendation topics and sentiment polarities and clustering similar reviews. By using their proposed approach, the online business sites and the managers of different organizations can find out the spam groups, those are uploading the spam information for better or worse rating about the products and take necessary action to control them**.**

3. Methodology Used

3.1.Sentiment analysis type

Based on the type of the output, Sentiment analysis can be carried out in different level. The Sentiment classification process can be of following types [12]:

- Binary: When the output of the system is considered as either positive or negative types, in such case the approach is called as binary as only two classes is considered. The positive class can be considered as accepted class and negative class can be considered as rejected class.
- Multi-class: When the output of the system is more than two classes, it is called as multi-class. This type of classification is mainly used, when the reviews are classified into different categories like the student’s grades, quality of movies etc.

In the present approach, the comments are preferred to be classified into either positive or negative class, i.e., the binary class is considered for analysis. The unlabelled comments are also analysed in the proposed work and thus, the reviews are categorized into two clusters for analysis.

3.2 Dataset considered for analysis

In the proposed approach, the dataset used for analysis must have both category of reviews i.e., unlabelled and labelled both. Thus, the IMDb movie review dataset is considered for analysis [13]. The detailed information about the IMDb dataset is provided as below:

- aclIMDb Dataset: The dataset considered for analysis should be available free for all the researchers and also used by most of the authors, so that the result can be easily verified with the result obtained. The aclIMDB dataset is freely available and used by authors working in the area of sentiment classification. The proposed work needs the reviews both in the form of labelled and unlabelled. Thus, the dataset contains both unlabelled and labelled reviews separated into training and testing categories. The dataset contains 12500 reviews both positive and negatively labelled along with this an unlabelled dataset having 50000 reviews are also present.

3.3 Machine learning approaches adopted

In this work, Support Vector Machine (SVM) techniques are considered for analysis. The detailed of the machine learning technique is as follows.

Support Vector Machine Classifier (SVM): The SVM technique is considered for analysis as this approach gives a better accuracy value in compared to other machine learning approaches [12]. It is observed that the machine learning techniques perform a better result in the area of sentiment classification and thus, it is preferred to other approaches.

In order to perform the classification task, SVM uses decision boundaries to separate the reviews into different categories by using the concept of hyperplanes. The hyperplanes are designed in such a manner that the separation between the two classes of values is as large as possible.The SVM technique uses the optimization principle on the training reviews which are represented in the form of a labelled pair like (x_j, y_j) , j= 1,2,3,…. where,

$$x_j \in uni^n andy_j \in \{1,-1\}^{\mathrm{k}} \qquad (1)$$

$$\alpha,\beta,\gamma \frac{1}{2}\, weight^T weight + Constant \sum_{k}^{j=1} \xi_i \; subject\ to\ y_j\left(weight^T \xi\left(x_j\right)+c\right) \geq 1-\gamma_j\,,\ \ \gamma_j \geq 0 \qquad (2)$$

The training vector x_j need to be mapped to a high dimensional space by x_j. As the SVM technique does not work on the reviews directly i.e., consider the input in the text format, it needs to be converted into a numerical form using different functions. After the reviews are converted into numerical form, scaling need to be performed to maintain a range in which the values fall in i.e., between [1,0].

3.4. Transformation of Reviews

The comments are written by the authors are in their native language. These comments cannot be processed by machine learning techniques and thus, they need to be transformed into a form i.e., accepted by the machine learning techniques for further processing. The different functions used for the transformation of reviews into suitable machine understandable form are as follows:

- CountVectorizer (CV): This function mainly concerned with occurrence of the features i.e., the words in the reviews [14]. Thus, the final matrix obtained using this approach is sparse in nature. The frequency of the features or words is not taken into account in this approach
- Term Frequency - Inverse Document Frequency (TF-IDF): The frequency of the word present in the reviews plays an important role [14]. The TF-IDF takes care about this situation. The TF is concerned

with the frequency of word in a review or comment and IDF concern with the frequency of word in whole dataset.

In the present work, TF-IDF preferred to CV for conversion as word frequency in a review plays a vital role while specifying the polarity of review.

3.5. Performance Evaluation Parameters

In order to check the performance of a machine learning approach, its performance must be verified using a proper mechanism. Confusion matrix is a matrix helps to check the performance of the machine learning techniques and this approach is universally accepted for performance evaluation of ML techniques.

This approach works on the labelled reviews mostly. The original label of the reviews is present and the reviews are predicted using ML techniques. If the original label of comment is positive and predicted one is also same, then it is marked as true positive (TP) or else it is marked as false positive (FP) [15]. Similarly, If the original label is negative and predicted one is also negative, then it is denoted as true negative (TN) or else it is marked as false negative (FN)

Confusion matrix also known as contingency table is typically used in supervised machine learning technique for the purpose of visualize the performance of algorithm. From classification point of view, True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FP) is used to compare label of classes [15]. True Positive represents the reviews that are positive also classified as positive by the classifier whereas False Positive are positive reviews classified as negative. Similarly, True Negative represents the reviews which are negative also classified as negative by the classifier where as false negative are negative reviews classified as positive.

The information obtained from the confusion matrix, suggest the follows parameters that helps to check the performance of the ML techniques

- Precision: It suggests the exactness in polarity predicted by classifier. This value is calculated for both categories of reviews. Precision is obtained by finding the ratio of correctly predicted reviews to the count of same polarity reviews shown as below:

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{Falsepositive}} \qquad (3)$$

- Recall: Completeness is an important criterion for any machine learning performance evaluation. Recall performs this task. Recall is obtained by ratio of correctly predicted reviews to reviews belonging to the same polarity group shown as below:

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \qquad (4)$$

- Accuracy: This parameter is used by most of the author to show the quality of their result. It is obtained by ration of correctly identified reviews to all the reviews present in the dataset, shown as below:

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{TruePositive} + \text{Falsepositive} + \text{TrueNegative} + \text{FalseNegative}} \qquad (5)$$

4. Proposed Approach

Present work proposed an aspect-based sentiment analysis of movie reviews. The IMDb movie review dataset is considered for analysis. The movie reviews are first filtered based on the keywords like romantic, horror, action etc. After the reviews are filtered based on a particular aspect, the pre-processing is done. During the pre-processing phase the words that do not have any effect to the sentiment of the reviews are removed, the numbers and the special characters are also removed. The reviews are mostly textual in nature and thus, for the purpose of machine learning they are transformed into matrix of numbers. Then given input to the machine learning for classification. For the classification purpose, Support Vector Machine (SVM) technique is used.

The process can be explained in detailed using following Figure 1:

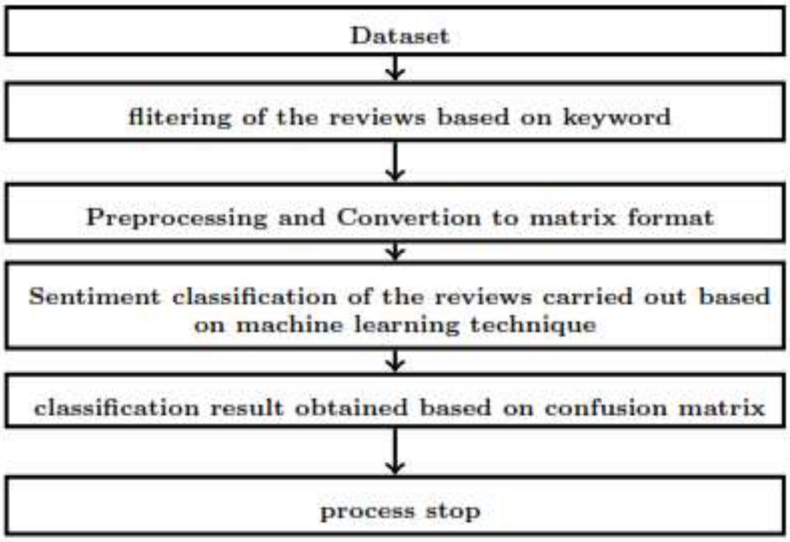
Dataset
flitering of the reviews based on keyword
Preprocessing and Convertion to matrix format
Sentiment classification of the reviews carried out based on machine learning technique
classification result obtained based on confusion matrix
process stop

Figure 1: Flow chart for proposed approach

The steps carried out to obtain the final result of the proposed approach can be explained as follows:

1. In the present work, aclIMDb dataset is taken into account for labelling the labelled reviews [13].
2. The movie reviews are filtered based on the keywords. In this paper, we have considered the romantic movies for classification and thus, the reviews containing the words like romance, love, romantic are considered for analysis. The reviews that do not contain these keywords are neglected for further analysis.
3. As the reviews are written by the customers in natural language, they may contain information that are not necessary and thus, need to be removed. This informationis
    - Stop words: Like other languages, English language also have words that are used in text often and just have any sentiment impact on the text such as the. Thus, the most frequent used word is found out and removed for better analysis.
    - Numeric and special character: The recent review writing style includes numbers and special character, having no effect on sentiment value and thus, removed.
    - Stemming and lowering the case: A particular word may be used in different forms in a text. Thus, while processing these words are considered as separate entities. In order to avoid the situation, all the words are transformed into one case i.e., either lower or capital [15]. In this case, the lowering of the words is considered. Again, the root words of each word are found out using stemming. So, there is a control on the number of words.
    - After removing unwanted information, uniform the case, and representing the words in root form, the next step is to transform the reviews into a numerical matrix. In this work, TF-IDF is used to transform the processed reviews into matrix form which is understandable by ML techniques.
4. After the transformation of the reviews both training and testing to ML understandable form, the training dataset is used for training purpose and the testing dataset is tested based on the learning information obtained from the training dataset. The IMDb dataset has 25000 reviews for training and testing. After the filtering of the reviews, the dataset size reduced from 25000 to 11658 reviews for training and similarly, the testing dataset size also reduced from 25000 to 10985. These reviews are transformed to matrix of numbers and given to SVM for classification. The result obtained after classification is shown in below Table 1.

Table 1: Parameter obtained using SVM classifier

Confusion Matrix			Evaluation Parameters		Accuracy
	Original Labels		Precision	Recall	89.884
	Positive	Negative			
Positive	9667	1318	0.88	0.91	
Negative	989	9996	0.87	0.93	

5. After the classification of the reviews are carried out, the result of the confusion matrix i.e., the accuracy is checked to know whether the result obtained is up to an acceptable standard on not.

5. **Conclusion and future work**

In the present work, the aspect-based sentiment analysis of the movie's reviews is carried out. The movie reviews are generalized in natured. These generalized reviews are not categorized into different categorizes like romance, horror, action, suspense etc. but the movie viewers prefer to watch a particular type of movies. Thus, this approach helps the viewer to know the categories or movies and based on that they prefer to watch the specific movie or not. This approach can be used in case of mobile reviews for different categories like camera quality, screen size, internet speed, display quality, storage space and many different factors. Thus, the analysis helps them to go for those mobiles, which matches their requirements and based on that they plan for future.

The writing style of the reviews varies from person to person. Many of them do not use the typical words like romantic, suspense and others, in such case, the synonym words also searched in the reviews for the better accuracy result. Again, there are cases, where the viewers may go for multiple keyword search i.e., romantic suspense movie in that case both the keywords combined together to obtained the better result.

References

[1] B. Liu, Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies, 2012.

[2] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, pp. 82-89, 2013.

[3] Z. Xiaojin and G. A. B., Introduction to Semi-Supervised Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning, 2019.

[4] G.Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in *Seventh International Conference on Contemporary Computing (IC3)*, 2019.

[5] T. Hastie, R. Tibshirani and J. Friedman, Unsupervised learning, Springer, 2009.

[6] M. F. A. Hady and F. Schwenker, Semi-supervised learning, Handbook on Neural Information Processing, Springer, 2013.

[7] S. Yadavilli and K. Seshadri., "A Framework for predicting item ratings based on Aspect Level Sentiment Analysis," in *International Conference on Advance Computing and InnovativeTechnologies in Engineering (ICACITE)*, 2021.

[8] L. Y, F. Y, L. Y and C. G. Z. A., "Aspect-based sentiment analysis for online reviews with hybrid attention networks.," *World Wide Web.*, vol. 2, pp. 1-9, 2021.

[9] Z. Y, L. H, J. C, L. X and T. X., "Aspect-Based Sentiment Analysis of User Reviews in 5G Networks," *IEEE Network*, vol. 35, no. 4, pp. 228-233, 2021.

[10] S. C, S. B and S. SP., "Aspect-Oriented Sentiment Classification using BiGRU-CNN model.," in *5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021.

[11] L. J, L. P, X. W, Y. L and Z. P., "Exploring groups of opinion spam using sentiment analysis guided by nominated topics.," *Expert Systems with Applications*, 2021.

[12] A. A. R. S. K. Tripathy. A., "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications,* vol. 57, pp. 117-126, 2016.

[13] A. L. Maas, R. Daly, P. T. Pham, D.Huang, A. Ng and C.~Potts, "Learning word vectors for sentiment analysis," in *49th Annual Meeting of the Association for Computational Linguistics: Human anguage Technologies-Volume 1. Association for Computational Linguistics*, 2011.

[14] R. Garreta and G. Moncecchi, Learning scikit-learn: machine learning in Python, Packt Publishing Ltd, 2013.

[15] Tripathy.A., Anand. A. and Rath. S. K, "Document-level sentiment classification using hybrid machine learning approach," *Knowledge and Information Systems,* vol. 53, no. 3, pp. 805-831, 2017.