International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)
ISSN (P): 2249-6890; ISSN (E): 2249-8001
Vol. 8, Issue 4, Aug 2018, 211-228
© TJPRC Pvt. Ltd.

ASPECT BASED OPINION PREDICTION USING DIVISIVE ANALYSIS FOR THE USER RECOMMENDATION SYSTEM

M. JOHN BASHA[1] & K. P. KALIYAMURTHIE[2]

[1] *Research Scholar, Department of Computer Science Engineering, Bharath University, Chennai, Tamil Nadu, India*

[2] *Professor & Dean, Department of Computer Science Engineering, Bharath University, Chennai, Tamil Nadu, India*

ABSTRACT

Social media is a vastly developing technology in the system environment on the internet and this knowledge assistance was useful for many organizations, people or company to make correct decisions about the products, things, and the recently released movie. Opinion mining is used to track the emotions of the public about a specific product and which is one kind of natural language processing or otherwise called as sentiment analysis. But, in case of large reviews about the product, the particular feedback prediction was the major drawbacks. To make a valuable decision about the manufactured product based on the proposed technique of Coherence-based Aspect Opinion Pairs Detection (CAOPD) framework. Initially, preprocessing the input dataset to remove the stop words and extract the relevant keywords (i.e noun, adverb, verb, an adjective based words). By using the Map Reduce (MR) methodology to perform parallel operations with reduce the size of the input data and speed up the system. Then, using the Divisive Analysis (DIANA) method based Nearest-Neighbor Clustering (NNC) algorithm to evaluate the distance and similarity between the extracted keywords and make clusters. This analysis is otherwise called a top-down approach and thus formed a set of active clusters and make the decision between the aspect and opinion of the customer reviews. To split individual reviews (i. e paragraph into the sentence) and applying Part of Speech (PoS) tagging method to extract the aspect and its opinions. Then, finding the coherence range between the aspect opinion pairs based on the Coherence-based Aspect Opinion calculation process. In this work, the fuel and engine recommendation also implemented for suggesting the best fuels and engines used in the machinery. Finally, estimate the relativity of the user review of the similar opinion and aspects word. Therefore, the proposed CAOPD method is compared with the various techniques such as CFACTS-R, FIFS, K-means (TF), K-means (PMI), DF-LDA, L-EM, and the PSM in terms of entropy, purity, precision, recall, accuracy, efficiency. Therefore, the proposed CAOPD system achieves greater performance than the other techniques.

KEYWORDS: Part of Speech (PoS), Aspect-Opinion pairs, Divisive Analysis & Nearest Neighbor

Received: May 11, 2018**; Accepted:** Jun 01, 2018**; Published:** Jul 04, 2018**; Paper Id.:** IJMPERDAUG201824

1. INTRODUCTION

Opinion Mining is a subfield of data mining that analyses a document with large collection of datasets. The collection and inspection of the opinions about the product that completed through the comments, tweets or reviews, blog posts which are employed to make a system of the building. The opinion mining or otherwise called as sentiment analysis[1, 2]. The public produces a comment about the product or theme based on their emotion is referred to as the sentiment analysis and which is name as a category of natural language processing. It helps to analyses the newly launched product or any other services. There are numerous ways of challenges presented in the sentimental analysis: positive reviews, negative reviews, neutral review, and partially based reviews. Figure 1 shows the different opinions among the persons.

Original Article

Figure 1: Opinion Mining

The various methods and the techniques are learned in this research for analyzing the product comments. There are 6 broad dimensions is reviewed such as lexicon creation, product aspect extraction, review usefulness measurement, sentiment classification, subjectivity classification, opinion word and also discussed numerous applications of opinion mining. The product feedbacks are collected and make the percentage of the score to sale the product effectively. In this manure, recommendation system or recommender systems [3] are introduced which is a subdivision of information filtering system. The appropriate rating or preference of an item (for example movies, books or music) delivered from the RS based on the user query response. There are three different types of RS such as content-based, hybrid based, and the collaborative filtering based recommendation. Thus, the existing recommendation system necessities the improvement for user requests for better recommendation qualities. The enhanced personalized location recommendation system [4] is desired for improving RS algorithm and user location preference model. Initially, recommend a hybrid user location preference model of sentimental analyze technique to extract the check-ins and text-based tips which are processed them to develop a “preference”. After that, both of the venue similarity influence and the user social influence are used to predict the accurate location based on the developed social location matrix factorization algorithm. Furthermore, improving the performance of location-based recommendation system. Hereafter, revise some existing techniques to extract different opinion and produce a perfect result. Therefore, recommend a pattern based approach [5] which is used in the online social networks like Twitter for multi-class sentimental analysis (SENTA). The wide set of features such as content and form which are extracting from the text by using the user-friendly tool SENTA. The recommendation system can also be used in a mechanical and production field, in which the recommended practice covers the load testing and determines the horsepower output of diesel engines. Also, a standard procedure is established within the practical limits for determining the performance of engines based on the following conditions: in which, the condition of equipment is determined and the equipment inspection reports are tested. The components have been considered in this system are brake horsepower, main generator, and traction horsepower. In a fuel recommendation system, the oil drain intervals can be extended based on the oil analysis, which reduces the potential risks of failures that associated with the oil drain periods. All maintenance requirements should be determined based on the operation and maintenance manual. In which, the maintenance interval schedule is utilized to illustrate the servicing intervals.

In most opinion mining applications, a pattern related features are utilized to improve the classification accuracy rate. The binary and ternary classification is classified into 7 different sentimental classes such as fun, happiness, neutral, love, hate, sadness, and anger respectively. The classification accuracy needs to improve in furthermore techniques. The probabilistic supervised joint aspect and sentiment model (SJASM) [6] is presented for identifying the similar opinions

and the opinion based emotions. This model can solve the one go problem further down a unified framework and predict the overall rating of aspects. Then, established a collapsed Gibbs sampling technique for the comprehensive inference of SJASM. Moreover, needs to enhance the prediction efficiency. The new topic model[7]for the complementary aspect-based opinion mining through the asymmetric assortments that suggests the Cross-collection Auto-labeled Max Ent LDA (CAMEL). All the conforming opinions are keeping at safe and modeling both the common and specific aspects of collections that gained the information in the CAMEL. The auto-labeling scheme (AME) is employed to distinguish among the aspects and opinions words without elaborative human labeling. After enhancing this scheme by utilizing additional word based similarity embedding is named out as a new feature. Then, the coupled Dirichlet Processes (DP) is additionally added with CAMEL to enhance the suggested scheme of a nonparametric alternative called as CAMEL-DP. Moreover, improve the efficiency of distinguishing words by using clustering based analysis.

Next, this work has made the following main objectives:

- To deliver a distributed data storage and investigate the support data-intensive distributed applications.
- To estimate the Opinion Mining on large e-commerce document datasets based on the product reviews.

The remaining portion of this article is prearranged as follows: The related works on the opinion mining is surveyed in Section II, describe the proposed Aspect-based Opinion Mining (AOM) in Section III, and derive the detailed experimental results of the proposed framework in Section IV. Section V, concludes the proposed technique.

2. RELATED WORK

This related work division deliberates the several techniques presented in the previous research work. The evaluation of traditional techniques capability in the opinion mining, that necessitates illustrating some drawbacks were explained in the following sections. [8]studied the opinion mining and the sentiment analysis to design new avenues. The people needed the opinion from the multiple options and the right time was employed to make a decision. Then, the sentimental analysis involved to choose the correct reviews among the valuable resources. These were the major description viewed in this work to predict the decisions. [9]described the concept-based opinion mining for enriching Sentic Net perceptions with disturbing data for allocating the sentimental tag. The Word Net-Affect (WNA) was most helpful to classify the subset of SenticNet concepts. The classification results were based on emotion-related corpus such as various psychological features and several concept similarity measures. Hence, the Support Vector Machine (SVM) classifier produced the high range of complexity error. [10] presented the Map-Reduce and Bulk synchronization Parallelism framework for reducing the large size data storage. In this work, studied the proficiency of BSP and MR framework, because BSP was more effective and fault tolerance than the MR framework. The K-Means clustering algorithm was utilized to handle the large size data set, but it occurred insufficient data storage problems. This problem has resolved and made sufficient storage space by utilized the BSP framework. [11]suggested the hybrid classification approach to feed the reviewers comments on the twitter based on the twitter opinion mining (TOM) framework. The main objective of this classification approach is to improve the accuracy rate and reduced the data sparsity problems. In a pre-processing step, the word was checked with the English dictionary and classified the tweets based on their emotions that utilized polarity classification algorithm(PCA). The positive, negative, and the neutral labeled tweet was the necessary classified output results. Moreover, the classification accuracy was improved in the upcoming methods.[12]presented a novel semantic based friend recommendation system for social networks which was named as Friend book.

The smartphone sensors discovered the user's lifestyle information that exploited from the Friend book. Initially, pre-processed the collected raw data on the smartphone utilizing a median filter, and extract the user's lifestyle based on the Latent Dirichlet Allocation (LDA) algorithm of the probabilistic topic model. After extracting the feature, construct a friend-matching graph to recommend the friends to users based on their matched life styles. Afterwards Friend book delivered the highest recommended based on the query user. Lastly, the recommendation accuracy of user's feedback was improved by incorporated a linear feedback mechanism. [13]presented a deep convolutional neural network to aspect extraction in the opinion mining. They utilized 7-layer DCNN such as one input layer, two convolution layers, two max-pool layers, a fully connected layer, and with softmax output layer utilized to segregate as aspect or non-aspect of the opinionated sentence. Finally, established a set of linguistic patterns for improving the classifier results. [14]recommended a two-phase unified model for simplifying the knowledge distribution and transmitting the context to social interaction. They illustrated the phases as,

- Initially, the various hidden relationships such as semantic relations, interactions, influences, temporal and spatial dependencies were extracted and uncovered from an enormous assemblage of plain-text-based context. As a result, the heterogeneous manufacturing networks data was produced through the semi-supervised algorithm.
- In the second phase, the recognized manufacturing network evidence data comprised many connection types and attributes that employed to match and gather network patterns. It comprises three main challenges such as high computational complexity, balance the larger data networks with thousands of nodes and edges, interactive semi-automatic matching was supported by the prototype system.

These were the major process performed in the social manufacturing paradigm. This conveying process was supported by different enterprise capabilities and incorporated the resources.[15]suggested the generic pattern based matching framework for addressing the problem of heterogeneous matching events with the pattern. They planned to increase the matching efficiency to adapt pruning with several bounds of matching scores. After that, introduced heuristic approach for distinguishing the NP-hardness of the optimal event matching problem with patterns. The efficiency was the major limitations and must be improved in the further process.[16] conducted an ex-situ tensile fatigue fracture test for estimating the temperature, humidity, and stress of the fuel cell membrane.

Here, the fatigue lifetime was measured based on the number of cycles. [17] utilized an Artificial Neural Network (ANN) technique for predicting various parameters used in a diesel engine. In this modeling, the emissions such as smoke level, Carbon monoxide (CO), Nitrogen Oxide (NO_x), Brake Specific Fuel Consumption (BSFC), and Brake Thermal Efficiency (BTE) have been considered. [18] analyzed the performance of diesel engine by comparing the use of mineral diesel and biodiesel obtained from the cottonseed oil. The measures considered in this evaluation were fuel consumption, brake specific fuel consumption, and brake thermal efficiency.

2.1. Motivation of Work

In the existing research work specifies the text pattern matching in the contemporary Intrusion Detection Systems (IDS) that delivered the guarantee of deterministic time. A Hierarchical Agglomerative Clustering (HAC) algorithm was performed to clusters the similar patterns based on the results from the similarity measure. But, it executed overlapping phenomenon that reduces the movement of the irrelevant pattern from one cluster to other cluster and it was performed under the principle of the bottom to top approach and then moves to the parent. In an early stage, it grouped the incorrect

data of the relocation of objects. The main drawbacks of this system produced high complexity and high computational time.

3. PROPOSED WORK

This section considers the implementation portions of the proposed Coherence based Aspect Opinion Pairs Detection (CAOPD) framework for evaluating the sentimental reviews of the products. This CAOPD framework is used to efficiently categorize the movie reviews based on the aspects and their opinions. This recommendation framework is also utilized for the machinery recommendation, in which the best fuel and engine have been recommended to the users based on its consumption and energy efficiency. Figure 2 shows the workflow of the proposed CAOPD framework.

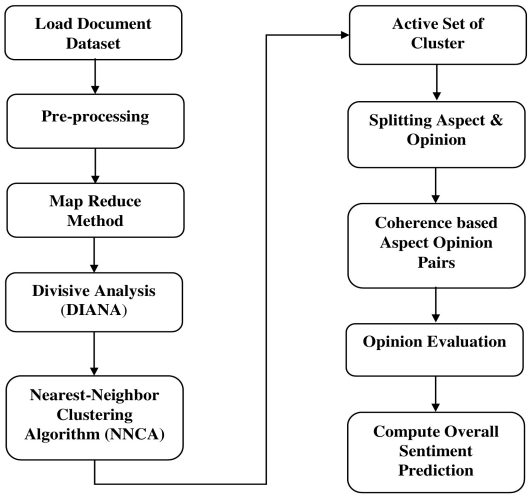


Figure 2: Workflow of Proposed Work

Load the movie document dataset as the input of the proposed CAOPD framework. First of all, clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies in the data preprocessing stage. Then, the extracted keywords are the required source of the Map Reduce (MR) technique to efficiently compress the data and speed up the system in a parallel manure. The large size of the dataset can successively be compressed in this MR technique. After that, measure the distance between the keywords and also evaluate the similarity to make a cluster. The user comments comprise both the aspect and the opinions are included in that cluster. Thus, the Nearest-Neighbor Clustering Algorithm (NNCA) is used to find the minimum distance between the clusters. This analysis is performed based on the neighbor discovery are called as Divisive Analysis (DIANA) method or top-down clustering approach. Therefore, the active set of clusters formed based on the set of aspect and the opinions of individual reviews. It can be identified and split the aspect and its corresponding opinions. After that, derive the coherence value among the aspect opinion pairs and evaluate the opinion. Finally, compute the overall sentimental score of a view based on their aspects and its opinion. Moreover, the measures such as brake thermal efficiency and fuel consumption are considered for recommending the machinery to the requested users. The fuels such as petrol, diesel, and gas are used in most of the machinery in which the petroleum is extensively used in vehicles. Based on the fuels used in the machines, the engine is categorized into the types of diesel engine, petrol engine, gas engine, and electric engine.

Table 1 presents the variables used in the proposed algorithm.

Table 1: Symbols and Descriptions

List of Symbols	Description
Sw_p	Senti Word Net polarity score
Wd_n	Set of words from all the reviews
n	Number of words
$U^{A,(k)}$	List of most probable words in aspect k
$U^{O,(k)}$	List of most probable words in opinion k

3.1. Pre-Processing

A Preprocessing stage is an important process to reduce the redundancy and improve the clustering algorithm efficiency. Therefore, it is necessary to preprocess the input movie dataset wisely. The main intention of stop-word removal (SWR) process is to obtain the key terms or key features from the online reviews text document and to improve the relevancy among word and category. Table 2 shows the list of stop word.

Table 2: Stop Words List

Stop words[19]	# , a, a's, able, about, above, according, across, all, actually, after, afterwards, again, ain't, against, allow, almost, alone, along, already, also, am, although, always, among, amongst, an, and, anything, causes, certain, certainly, changes, clearly, being, believe, below, beside, besides, best, better, least, less, lest, let, ourselves, out, outside, over, that's, the, their, theirs, under, unfortunately, unless, unlikely, until whoever, whole, whom, whose, why, yours, yourself, yourselves, etc.

The part of the speech (PoS) tagging procedure is used to break each document into sentences and then utilized the Stanford parser to extricate just noun, verb, adverb, and the adjective expressions and then evacuate the non-word tokens, for example, numbers, HTML labels, and accentuation. In this situations, consider just a noun and verbs since it has comprised 80% of the aggregate terms of a review. A large portion of the words utilized as a part of English parlance is called as pointless words (more than 400 words presence) that are to be extricated by utilizing the content mining or information Recovery (IR) technique. These useless words are called 'Stop words' that convey no data (i.e., pronouns, relational words, and conjunctions). Therefore, the stop words are removed in the pre-processing stage and are demonstrated as imperative for predicting the overall sentimental scores.

3.2. Map Reduce

The Map-Reduce MR method is used to partitions the large database in a parallel way where the individual employments are prepared by the mapping process and after that reducing the overall process to yields the reduced results. The extracted keywords are considered as the input of the system and map the words. After that, the allocated keyword are compressed to reduce the size of the data. Then, recalculated the keywords and reduce the redundant words based on the reducer. Figure 3 represents the MR methodology framework.

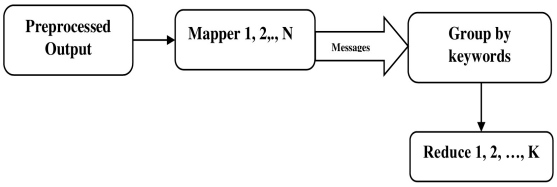


Figure 3: Map Reduce Methodology

3.3. Divisive Analysis Based Clustering Algorithm

The proposed divisive analysis based nearest neighbor clustering algorithm procedure is performed based on the following calculations as:

3.3.1. Algorithmic Formulation

The proposed computational method is performed based on the publically available library Senti Word Net[20]. In this library, the lookup table has identified the sentiment score of the selected reviews of the text. From this lexical resource, the individual term t comprising in the WordNet which accompanying into three numerical scores such as positive, neutral, and the negative in terms of pos(t), neu(t), and the neg(t). These three terms are used to extract the relevant opinionated terms and their corresponding scores in the Senti Word Net. In this library, the linguistic features are used to take a lot of decisions, predicted an individual linguistic features weights, and the associated sentiment scores of the aggregation methods are all produced. Hence, the prediction of overall sentiment polarity score of the algorithm is based on the implementation of Senti Word Net.

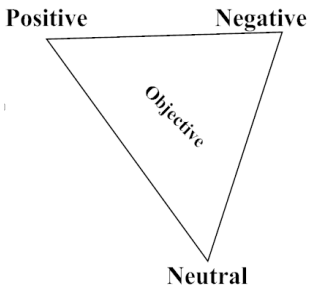


Figure 4: SentiWordNet

3.3.2. DIANA Based Nearest Neighbor Clustering

The divisive hierarchical algorithm or divisive analysis (DIANA) which is used to build the hierarchy based on the top-down approach. This proposed algorithm is an inverse order of the Agglomerative Hierarchical Clustering algorithm. There is totally *n* number of words that comprising in a single large cluster. Consequently, an individual step represents that the largest available cluster is split into the small number of the clusters till to reach all the clusters contains a large set of keyword and then satisfied the termination condition.

Where $n-1$ step construction process is assigned in this hierarchy manner. In this DIANA, all possible fusions of two objects are considered leading to 2^{n-1} probabilities which split the reviews into two clusters. Whereas, the previous agglomerative method indicates only $n(n-1)/2$combinations. A Nearest Neighbor Clustering (NNC) algorithm is utilized to group the similar keywords based on the minimum distance with the help of cosine and Jaccard similarity measurement.

The possible number of clusters is considerably greater than the previous agglomerative method. In order to overcome these large computational problems, to follow the steps discussed below:

The DIANA follows the top-down approach that assigns the single cluster having the hierarchy level and the arrangement number as,

$$HL(0) = n; \quad (1)$$

$$q = 0; \quad (2)$$

Then, derive the most nearest neighbor keywords based on the distance and similarity measurement of cosine and jaccard calculation. First of all estimate the jaccard similarity measurement as follows. To compute the length between the length of sentence1 and sentence 2 as,

```
int length= Math.max(sentence1.length, sentence2.length);
```

Then set the array of sentence as,

```
sentence1 = java.util.Arrays.copyOf (sentence1, length);
sentence2 = java.util.Arrays.copyOf (sentence2, length);
```

From the array of extracted keywords, the row and column wise calculation is implemented. Initialize the inter and union as,

```
int inter = 0;
int union = 0;
for (int i = 0; i < length; i++)
{
if (sentence1[i] > 0 || sentence2[i] > 0)
        {
union++;
if (sentence1[i] > 0 && sentence2[i] > 0)
        {
inter++;
}} }
return (double) inter / union;
   }
public double distance(String s1, String s2) {
return 1.0 - similarity(s1, s2);
```

Then, derive the similarity distance between the keywords based on the cosine calculation is as follows:

```
int length = Math.max(sentenc1.length, sentence2.length);
    sentenc1= java.util.Arrays.copyOf(sentenc1, length);
    sentenc2= java.util.Arrays.copyOf(sentenc2, length);
doubleagg = 0;
for (int i = 0; i < length; i++)
    {
agg += sentenc1 [i] * sentence2 [i];
    }
returnagg;
  }
public double distance(String s1, String s2) {
return 1.0 - similarity(s1, s2);
```

The distance-based similarity calculation output creates the clusters with similar keywords. Finally, this calculation produces the single-single extracted keyword is created in all the clusters. Then initialize the SentiWordNet library with the corresponding polarity score value and take the set of words from all the reviews. The redundant words are removed based on the MR method to map and reduce the redundant words correspondingly. For estimating the sentiment score of X based on the removed redundancy keyword with the score value is updated and stop the process. Afterward, initialize the active set of clusters such as:

Table III

Strong Positive	C1
Positive	C2
Neutral	C3
Negative	C4
Strong Negative	C5

The proposed divisive analysis algorithm is listed as follows:

Divisive Analysis
Input: Input cluster (considering reduced keyword extracted data as a single cluster) **Output:** Active Set of clusters
Begin Initialize Sw_p **Step 1:** Let taken Wd_n **Step 2:** $Rd_n \leftarrow (Wd_n, Map)$ **Step 3:** for X ← 1 to N **Step 4:** $S_X = Rd_x(Sw_p)$; **Step 5:** End for X; **Step 6:** Initialize theClusters (C1, C2, C3, C4, and C5) **Step 7:** For y ← 1 to N **Step 8:** if (S_y>=1.0) C1 ← include (Rd_y); **Step 9:** Else if (S_y>0.0 && S_y<1.0) C2 ← include (Rd_y) **Step 10:** Else if (S_y=0.0) C3 ← include (Rd_y) **Step 11:** Else if (S_y <0.0 & $S_y \geq (-0.1)$) C4 ← include (Rd_y) **Step 12:** Else C5 ← include (Rd_y) **Step 13:** end for y **Step 14:** End

The proposed algorithm executes the set of active clusters based on the calculation process. If the similarity measure is greater than or equal to 1.0 as C1 (OR) greater than 0.0 and lesser than 1.0 as C2 (OR) equal to 0.0 as C3 (OR) lesser than 0.0 and greater than or equal to -0.1 as C4 (OR) else C5.

3.4. Coherence based Aspect Opinion Pair Prediction Framework

The proposed CAOPD framework is primarily used to predict the overall sentimental scores about the movie based on the reviewer's aspects and its opinions. In an opinionated text, the corresponding opinion and the aspects are detected and extract the relevant keywords and which represent a subdivision process of sentimental analysis. For distinguishing the particular parts of a film, the sentiment holder is either adulating or complaining. Additionally, the keywords indicate the movie aspects are as, 'animation', 'direct', 'music', 'sound effects', and 'scene' are all split and stored in the movie aspect feature list. These are some aspects are collected from the own dataset of online reviews. From this dataset, the paragraph or sentence covering an individual movie reviews $R$and manually splitting each reviewfrom the paragraph into the sentences.

After performing the annotation rule, to tag the keywords that matched the ones in the movie aspects and its opinions. For using the part-of-speech (POS) tagging method, to extract the aspects and its opinions of each sentence. For example, the verb section 'act' in the rundown coordinates any type of the verb 'act' (e.g. 'acts' or 'acted'), yet does not coordinate the noun term 'act'. Also, the noun passage 'sound' matches either singular or plural noun (e.g. 'sounds') yet not the verb term 'sound'. The longest coordinating strategy is connected, and every section in the element records is given an interesting comment code and utilized the POS labeling, the Stanford Log-direct Grammatical form Tagger for tagging the keywords [21]. Then, the coherence value calculation process is performed between the sentences i and j is denoted as

$$C_{i,j} = \rho \,.\, \varphi(i,j) \tag{3}$$

Where, ρ represents the weight vector, $\varphi(i,j)$ represents the similarity between the sentences and i,j represents the two input sentences. The semantic similarity degree is estimated between the high probabilities words through the single word distribution is called as coherence score measure value. If reached a higher score that represents the better quality.

The above-mentioned equation (3) can be used to predict the quality of an opinion. Then establish the coherence value among the one aspect and its corresponding opinions. Therefore, evaluate the coherence based on the high probability words T. The mathematical evaluation is as follows,

$$\mathrm{Coh}_{A,O}\left(k; U^{A,(k)}, U^{O,(k)}\right) = \sum_{t=1}^{T} \quad \sum_{l=1}^{T} \quad \log \frac{\mathrm{D}\left(u_t^{\;A,(k)}, u_l^{\;O,(k)}\right)}{\mathrm{D}\left(u_t^{\;A,(k)}\right)} \tag{4}$$

Where, $U^{(k)} = u_1^{(k)}, \ldots, u_T^{(k)}$ represents the list of T the most possible words of the opinion $k, D(u)$ represents the total number of documents comprising the word v , $D(u,u')$ represents the total number of documents comprising both the u and u', and ω represents the smoothing variable. In a document, observe the opinion word $U^{O,(k)}$ and previously detected the aspect word $U^{A,(k)}$ is used to evaluate the probability of

$$\frac{\mathrm{D}\left(u_t^{\;A,(k)}, u_l^{\;O,(k)}\right)}{\mathrm{D}\left(u_t^{\;A,(k)}\right)} \tag{5}$$

Thus, finally computed the coherence of Aspect-Opinion pairs. The proposed CAOPD algorithms is as follows:

Coherence based Aspect-Opinion pairs
***Input:** Individual Movie Reviews R, n* ***Output:** Coherence based Aspect-Opinion Pairs*
Begin *For each review* $R = \{r_1, r_2, \ldots, r_n\}$ ***Step 1:** Load the list of reviews in the Review List r* *For* $i = 1{:}i$ *is the review of the User with user id* U_i. ***Step 2:** Splitting each reviews* ***Step 3:** For each sentences we have applying PoS Tagging method, to extract Aspects and its Opinions.* ***Step 4:** Compute the Coherence of Aspect-Opinion pair using the equation (10)* ***Step 5:** End*

3.5. Sentimental Score Prediction

The main aim of this sentimental prediction is to analyze the sentimental score towards an individual aspect and its opinion of a movie review. For example, if a review can display one aspect or multiple reviews shows the similar aspects, or else multiple review represents the different aspects. In that case, calculate the overall aspects and the opinion about the movie and then split it. Thus, in the wake of computing the general feeling scores at sentence-level and deciding their audit perspectives in a sentence, the sentiment score for each survey review is ascertained by gathering together a similar viewpoint sentiment matches and taking the normal score. At that point, the sentence-level feeling score is figured by averaging the sentimental scores for all perspectives (i.e. both positive and negative scores). Thus, the conclusion score of the sentence 'this motion picture is great' (one positive sentence about general perspective) is more positive than the notion score of the sentence 'the film is great yet I don't care for the music' (an extra negative proviso about the music angle). Hence the overall sentimental score value is predicted based on the sentence level coherence estimation method and estimates the relativity of the user could observe the same opinion word and aspect word.

4. RESULTS & EVALUATION

This section use the Java tool to develop the aspect based opinion prediction and represents the effectiveness of the proposed framework by comparing with the existing FACTS, FACTS-R, JST, CFACTS, CFACTS-R, FIFS, K-means (TF), K-means (PMI), LDA, L-LDA, DF-LDA, L-EM, and the PSM in terms of words accuracy, precision, recall, accuracy, entropy, purity, coherent topics, and the clustering efficiency.

4.1. IMDb Dataset

In this research work, the input movie dataset is collectedfrom the IMDb [22] database and this proposed technique was implemented by using the Java tool to predict the sentimental score about the product. The customer can use the data for the commercial and non-commercial purposes and are located in the AWS S3 bucket named as IMDb-datasets. Daily can be refreshed the dataset and predict the correct judgment. The dataset comprises individual folder of each data reviews that includes tab-separated-values (TSV) format, and gzipped format.

4.2. Accuracy Vs Review

The accuracy is defined as the closeness of the standard measured value in terms of predicting the extracted keywords from the own created movie dataset. The various existing techniques FACeT and Sentiment extraction (FACTS), FACeT and Sentiment extraction with Rating (FACTS-R), Joint Sentiment Topic model (JST), Coherence based FACTS (CFACTS), CFACTS with Rating (CFACTS-R) [23] are compared with the proposed CAOPD framework. Figure 5 shows the accuracy versus review.

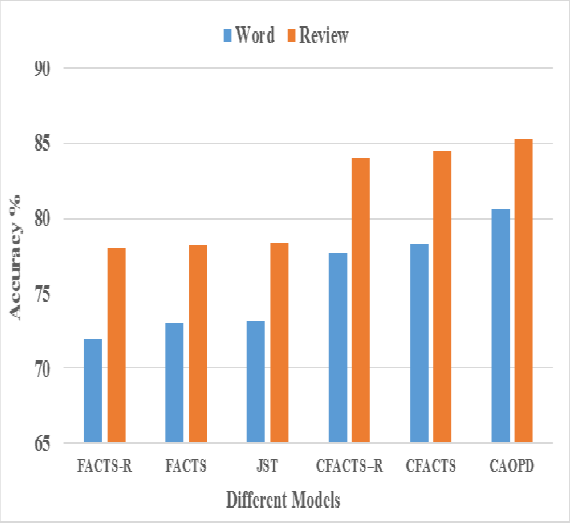


Figure 5: Accuracy Vs Words and Review

From the figure, clearly, represent that the proposed CAOPD technique achieves greater accuracy than the existing techniques. The word and the review extraction accuracy rate are greatly improved than the traditional. It approximately improves 2% and 1% than the existing CFACTS technique. Hence, the proposed technique outperforms than other.

4.3. Aspect Opinion Pair Vs Accuracy

The aspects opinion pair's accuracy rate is derived with the help of precision and recall measures. The precision is otherwise named out as the positive predictive rate which is the division of retrieved keywords that are relevant.

$$PR = \left\{\frac{\{aspect\ opinion\ based\ keyword\ \cap movie\ dataset\}}{movie\ dataset}\right\} \tag{6}$$

The recall is defined as the aspect of opinion based keyword extraction which represents the fraction of the extracted keyword to the original dataset that is effectively recovered.

$$Re = \left\{\frac{\{aspect\ opinion\ based\ keyword\ \cap movie\ dataset\}}{movie\ dataset}\right\} \tag{7}$$

The aspect opinion pair accuracy value is calculated with the help of precision and the recall range. The overall sentimental prediction capability is represented in terms of accuracy. The proposed CAOPD framework is compared with the existing techniques as Frequent item sets based Facet/Sentiment Miner (FIFS), JST, FACTS, FACTS-R, CFACTS, and the CFACTS-R[23].

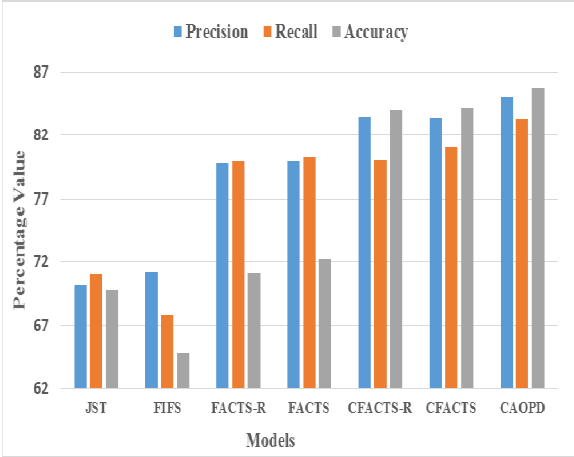


Figure 6: Aspect Opinion Pair Accuracy Rate

From the figure, shows that the accuracy, recall, and the precision range of the proposed technique is compared with the existing techniques. Thus, the proposed technique achieves more 17.5% of precision, 18.5% of recall, and 24.5%of accuracy than the existing techniques. Hence, the proposed CAOPD framework outperforms than the traditional systems.

4.4. Entropy and Purity

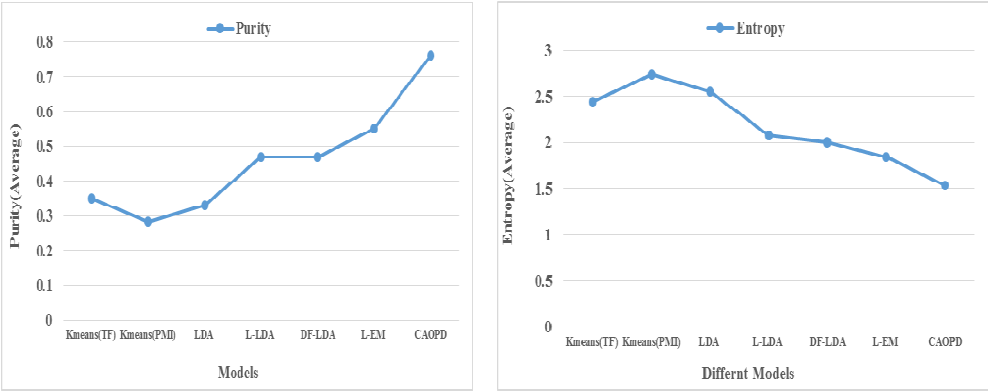
The entropy is the measure of predicting the clustering quality in terms of measures. In that case, the lesser entropy refers to the enhanced cluster quality and the superior entropy value refers that obtained low quality of clusters. Hence, the quantity of disorder is found by using entropy calculation. The mathematical formula of entropy as,

$$Entropy = \sum_i q_i\left(\sum_j\left(q_{ij}/q_i\right)\log\left(q_{ij}/q_i\right)\right) \tag{8}$$

Where, q_{ij}, q_i, q_j represents the extracted keywords from the sentences.

The purity is one of the principal authentication parameters for predicting the cluster quality. In that case, the evaluation of the cluster quality is compared with the original movie dataset is called as the purity. It is calculated as,

$$Purity = \sum_i q_i\left(max_j\left(q_{ij}/q_i\right)\right) \tag{9}$$



(a) **(b)**

Figure 7: (a) Entropy Rate, (b) Purity Rate

From the figure (a) and (b) shows that the proposed CAOPD framework is compared with the existing techniques such as: K means with term frequency (K-means TF), K-means Point wise Mutual Information (K-means PMI), Latent Dirichlet Allocation (LDA), L-LDA, DF-LDA, and the L-EM [24]. The proposed CAOPD framework achieves lower entropy and higher purity range than the existing techniques. Hence, the proposed works maximize the clustering efficiency.

4.5. Clustering Efficiency

The efficiency is defined as the effectiveness of the clustering algorithm in which cluster the relevant keyword or aspects to a similar group. The proposed CAOPD framework clustering algorithm is compared with the various clustering algorithms such as K-means TF, K-means PMI, Lexicon based, LDA, FCA, and the FFCA [25]. Figure 8 represents the clustering efficiency.

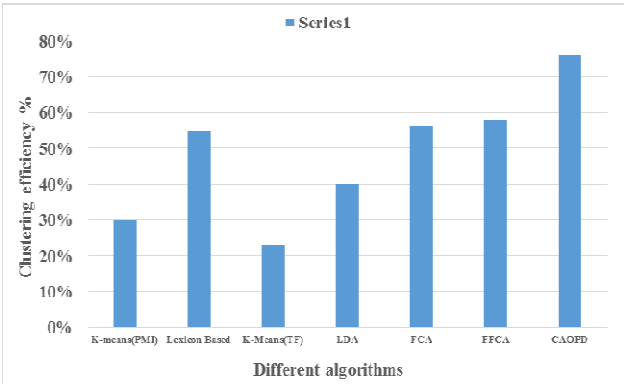


Figure 8: Clustering Efficiency

From the figure shows, the proposed clustering algorithm efficiency is compared with the existing techniques. The proposed clustering algorithm produces 76% efficiency than the other techniques. Hence, it reaches 23.6% more efficient than the existing FFCA algorithm.

4.6. Coherence Topics

The topic coherence is derived as the measure of extrinsic measure for unique client identifier (UCI) and the intrinsic measure for the University of Massachusetts Amherst (UMass) and both measures referred as the similar high - level idea. It can be calculated as,

$$Coherence = \sum_{i<j} sentimental\ score(w_i, w_j) \quad (10)$$

Where, $w_1, w_2, \ldots\ldots, w_n$ represents the extracted keywords.

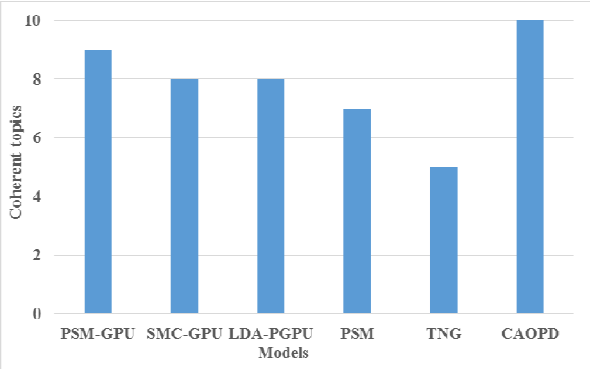


Figure 9: Coherent Topics

From the figure, shows that the proposed CAOPD framework coherence is compared with the existing techniques: Phrase Sentiment Model with Generalized Polya urn (PSM-GPU), semi-Markov CRF with GPU (SMC-GPU), (LDA-P-GPU), Phrase Sentiment Model (PSM), and the Topical N-Gram(TNG)[26]. The coherent topic range achieves higher than the overall existing techniques. Hence, the proposed CAOPD framework attains better than the existing techniques.

5. CONCLUSIONS

This paper presented a coherent based aspect opinion pair detection (CAOPD) framework for efficiently providing the feedbacks about the movie. In this paper, proposed the stop word removal technique to extract the relevant keyword and remove the unwanted words present in the movie dataset and using the Map Reduce (MR) technique to parallelizing the system vastly. Then, adopting the Divisive Analysis (DIANA) method to effectively investigate the large dataset and cluster the relevant keywords based on their customer online opinions using Nearest-Neighbor Clustering (NNC) algorithm. Then, estimate the coherence value among the aspect and the opinion using coherence based aspect opinion pair algorithm and evaluate the overall sentimental analysis about the movie result. The proposed CAOPD technique is compared with various existing techniques and produces more accurate results.

REFERENCES

1. *G. Vinodhini And R. Chandrasekaran, "Sentiment Analysis And Opinion Mining: A Survey," International Journal, Vol. 2, Pp. 282-292, 2012.*
2. *K. Ravi And V. Ravi, "A Survey On Opinion Mining And Sentiment Analysis: Tasks, Approaches And Applications," Knowledge-Based Systems, Vol. 89, Pp. 14-46, 2015.*
3. *L. Sharma And A. Gera, "A Survey Of Recommendation System: Research Challenges," International Journal Of Engineering Trends And Technology (Ijett), Vol. 4, Pp. 1989-1992, 2013.*
4. *D. Yang, Et Al., "A Sentiment-Enhanced Personalized Location Recommendation System," In Proceedings Of The 24th Acm Conference On Hypertext And Social Media, 2013, Pp. 119-128.*
5. *M. Bouazizi And T. Ohtsuki, "A Pattern-Based Approach For Multi-Class Sentiment Analysis In Twitter," Ieee Access, 2017.*

6. *Z. Hai, Et Al., "Analyzing Sentiments In One Go: A Supervised Joint Topic Modeling Approach," Ieee Transactions On Knowledge And Data Engineering, Vol. 29, Pp. 1172-1185, 2017.*
7. *Y. Zuo, Et Al., "Complementary Aspect-Based Opinion Mining," Ieee Transactions On Knowledge And Data Engineering, 2017.*
8. *E. Cambria, Et Al., "New Avenues In Opinion Mining And Sentiment Analysis," Ieee Intelligent Systems, Vol. 28, Pp. 15-21, 2013.*
9. *S. Poria, Et Al., "Enhanced Senticnet With Affective Labels For Concept-Based Opinion Mining," Ieee Intelligent Systems, Vol. 28, Pp. 31-38, 2013.*
10. *A. A. Golghate And S. W. Shende, "Parallel K-Means Clustering Based On Hadoop And Hama," International Journal Of Computing And Technology, Vol. 1, 2014.*
11. *F. H. Khan, Et Al., "Tom: Twitter Opinion Mining Framework Using Hybrid Classification Scheme," Decision Support Systems, Vol. 57, Pp. 245-257, 2014.*
12. *Z. Wang, Et Al., "Friendbook: A Semantic-Based Friend Recommendation System For Social Networks," Ieee Transactions On Mobile Computing, Vol. 14, Pp. 538-551, 2015.*
13. *S. Poria, Et Al., "Aspect Extraction For Opinion Mining With A Deep Convolutional Neural Network," Knowledge-Based Systems, Vol. 108, Pp. 42-49, 2016.*
14. *J. Leng And P. Jiang, "Mining And Matching Relationships From Interaction Contexts In A Social Manufacturing Paradigm," Ieee Transactions On Systems, Man, And Cybernetics: Systems, Vol. 47, Pp. 276-288, 2017.*
15. *S. Song, Et Al., "Matching Heterogeneous Events With Patterns," Ieee Transactions On Knowledge And Data Engineering, 2017.*
16. *Murthy, SV Niranjana. "Analysis & Prediction of Mischievous Behavior of Vehicle using ANPR and DBSCAN."*
17. *R. M. Khorasany, Et Al., "Mechanical Degradation Of Fuel Cell Membranes Under Fatigue Fracture Tests," Journal Of Power Sources, Vol. 274, Pp. 1208-1216, 2015.*
18. *N. Acharya, Et Al., "An Artificial Neural Network Model For A Diesel Engine Fuelled With Mahua Biodiesel," Singapore, 2017, Pp. 193-201.*
19. *E. M. Shahid And Y. Jamal, "Performance Evaluation Of A Diesel Engine Using Biodiesel," Pakistan Journal Of Engineering And Applied Sciences, 2016.*
20. *G. S. A. C. Buckley, "Smart Information Retrieval System At Cornell University."*
21. *Kamat, Vaishnavi. "Selecting base stocks for stock price prediction using minimum spanning tree and mahalanobis distance criteria."*
22. *S. Baccianella, Et Al., "Sentiwordnet 3.0: An Enhanced Lexical Resource For Sentiment Analysis And Opinion Mining," In Lrec, 2010, Pp. 2200-2204.*
23. *K. Toutanova And C. D. Manning, "Enriching The Knowledge Sources Used In A Maximum Entropy Part-Of-Speech Tagger," In Proceedings Of The 2000 Joint Sigdat Conference On Empirical Methods In Natural Language Processing And Very Large Corpora: Held In Conjunction With The 38th Annual Meeting Of The Association For Computational Linguistics-Volume 13, 2000, Pp. 63-70.*
24. *I. Datasets.*

25. *S. Mukherjee And P. Bhattacharyya, "Feature Specific Sentiment Analysis For Product Reviews," Computational Linguistics And Intelligent Text Processing, Pp. 475-487, 2012.*

26. *L. Chen, Et Al., "Clustering For Simultaneous Extraction Of Aspects And Features From Reviews," In Hlt-Naacl, 2016, Pp. 789-799.*

27. *W. Medhat, Et Al., "Sentiment Analysis Algorithms And Applications: A Survey," Ain Shams Engineering Journal, Vol. 5, Pp. 1093-1113, 2014.*

28. *A. Laddha And A. Mukherjee, "Extracting Aspect Specific Opinion Expressions," In Emnlp, 2016, Pp. 627-637.*