I.J. Information Technology and Computer Science, **2016, 3, 33-42**
Published Online March 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2016.03.04

Multi-Feature Segmentation and Cluster based Approach for Product Feature Categorization

Bharat Singh
School of Computing Science and Engineering, Galgotias University, India
E-mail: glabharat.mca@gmail.com

Saroj Kushwah and Sanjoy Das
GLA University, India, School of Computing Science and Engineering, Galgotias University, India
E-mail: {sarojkushwahsiem, sdas.jnu}@gmail.com

Abstract—At a recent time, the web has become a valuable source of online consumer review however as the number of reviews is growing in high speed. It is infeasible for user to read all reviews to make a valuable or satisfying decision because the same features, people can write it contrary words or phrases. To produce a useful summary of domain synonyms words and phrase, need to be a group into same feature group. We focus on feature-based opinion mining problem and this paper mainly studies feature based product categorization from the number of users - generated review available on the different website. First, a multi-feature segmentation method is proposed which segment multi-feature review sentences into the single feature unit. Second part of speech dictionary and context information is used to consider the irrelevant feature identification, sentiment words are used to identify the polarity of feature and finally an unsupervised clustering based product feature categorization method is proposed. Clustering is unsupervised machine learning approach that groups feature that have a high degree of similarity in a same cluster. The proposed approach provides satisfactory results and can achieve 100% average precision for clustering based product feature categorization task. This approach can be applicable to different product.

Index Terms—Product feature categorization, irrelevant feature, opinion mining, sentiment orientation, feature, cluster.

I. Introduction

Of late with invent of the internet technology, information technology has increased very fast and it has become more reliable than any other source of information. It's reliability and easy access prompting people to use more and more in their day to day life. Social networking twitter and blogs are connecting more and more people with each –other and sharing information globally.

This takes people very comfortable and sharing their opinion, experience and feedback regarding any product or entity the experience or feedback is very useful as well as very important for the manufacture organizations as they sell and their product based on the feedback and success to tailor their products as per demand and necessity of people. In the same way, monitoring consumer reviews is getting harder for the manufacturer and provider these need have inspired new way of research on extracting customer review.

Some researcher mainly focuses on the extracting the objective information from the review sentences in the document level opinion mining. Many recent studies on opinion analysis aimed to analyze and extract opinions of sentiment information and present them in the form of sentiment based or opinion-oriented summarization [12] [13].

In [14, 19, 21] only focus on the sentiment of the document and [10, 11, 24] to evaluate the orientation of reviews by using the supervised document classification algorithms. In [12] using the association rule mining for access the noun as frequent feature recognition. In [20] uses the web PMI point-wise mutual information to extracting the feature.

Many approaches have been proposed for building the aspect based opinion mining system. Almost each of them has some advantages and limitation as well.

While for aspect based opinion mining, the main task is to extract the aspect and the corresponding from the user review. In [1-9] had studied the problem. The proposed method uses the opinion word and context information to extract the irrelevant and relevant feature.

To identify the frequent and infrequent feature is a big problem. The problem is more complicated when extract the implicit feature. Different method is available to solve such problems addressed in [3,7].

But in the phrases level opinion mining, these opinions and information regarding any product are very useful for other people who want to buy the same product but it becomes difficult for the potential customer to analyze hundreds of reviews regarding the same products opinion mining approach are practiced to explore opinions to reach on final reviews of customer.

People use different words to represent the same aspect. For example connectivity: relation, communication, connection, transfer rate. Display: screen, touch pad. Music player: audio, playlist, song. Battery: cell, backup,

charging, long lasting etc.

The latest work done on aspect based opinion mining is opinion mining based on feature level. Methods considered are explicit and implicit feature opinion words are divided into two categories vague and clear opinion word to identify the implicit feature. The limitations associated with existing problems are required to be addressed and resolved in order to achieve relevant reviews with higher precision and accuracy.

To address these challenges, this paper explores the problem of feature- based opinion mining problem. The key contribution extended in three ways. First multi-feature segment method to segment the multi-feature sentence into the single feature unit.

Second implicit feature identification by using the context information of the sentence and the part of speech dictionary. The use of sentiment dictionary expansion, since an expansion of feature and sentiment based dictionary through an adjective, adverb, noun, thesaurus and word net tool.

Third categorization of feature of an entity or the product through or using the clustering, thus gives the best summary of an entity to users. Moreover, this will improve the performance of aspect based opinion mining system in terms of Average precision and accuracy.

II. Related Works

Different types of approaches have been proposed for aspect based opinion mining. We have discussed few of them in the subsequent section. In order to enhance efficient clustering based product feature categorization. It is necessary to address each problem associated with their approaches.

S. Momtazi et al., [1] widely discussed role of TREC (style give question answering system). The method proposed in [1] is a combined approach for expansion queries while considering two aspects; extracting more relevant data and finding more opinionative data. One of the method is select feedback term and opinion bearing term for querying expansion based on a chi-squired test and use the same query expansion to combine it in a linear existing scheme with the original query from the web. Y. Choi et al., in [2] describe techniques for sentiment analysis, which can be used for exploit sentiment topic information to generate context driven feature moreover the author analysis the domain-specific sentiment. The domain-specific sentiment analysis includes different steps their corpus representation. They have prepared a domain corpus which contains relevant parameter a set of query in the domain and they then combine it. Bootstrapping algorithm, sentiment topic which are retrieved from a sentence and contextually related to sentiment clues. The experimental results show that the bootstrapping algorithm is able to commerce and aggregate new clues. Also, verifies that the obtained domain context feature is more effective than generally used feature in sentiment analysis by running them on the same sentiment classified. Additional clue generated by bootstrapping algorithm that does not affect sentiment classification. Size of domain corpus does not affect the increased performance. W. J. Jia et al., in [3] describe the feature categorization of any product based on twice clustering method with semantic association. They focused on the use of opinion word in the place of context word to assess the interrelationship of the product features. Group information of opinion word received through automatically. The cluster results of active product feature are used as constraints, to cluster whole feature. For better performance, the twice clustering strategy has been preferred to single clustering. The experiment has been conducted at the initial stage and this approach has not been experimented with different language. C. L. Ferm ń in [4] described a domain specific resource based approach. They focus on describing the resources that capture the domain knowledge and proposes the extraction information to enable the induction of lexical-syntactic knowledge from an annotated corpus. The author use dependency recourses is a kind of list of pattern connecting feature words with opinion words, in order to compute accumulated precision and recall value. Only feature contained in domain taxonomy will be considered only well define subset. Large set of opinions, containing almost all the existing opinions but including a high number of non-existent ones.

J. Zhu et al. in [5] described a method for customer opinion polling from free from textual customer review, without requiring designing a set of question or assigning any rating. Firstly a multi-aspect bootstrapping method is proposed to learn the most valuable ART of each aspect that is used to identify the aspect. Secondly, the aspect based segmentation model is proposed to divide a multi-aspect sentence into a multiple single units for opinion polling, then author generates the opinion poll.

The experiment which was conducted in china on a real restaurant reviews is capability to demonstrate that this approach can be able to achieve 75% accuracy in aspect based opinion polling task. The proposed method that implies opinion polling does not need labeled training data, that's why it is very easy to implement and can an applicable to other languages.

In [22] authors proposed a data classification method multi-group discriminant linear programming (MDLP) for classification problems with Support Vector Machine (SVM), Neural Networks. Local optima problem addressed and shows the MDLP not eliminate this problem and solution is less complex. In result analysis small and medium sized market data is considered. Further, fuzzy delphi method used to select and gather data. The results show that MDLP is better than other existing methods for correct classification.

In [23] authors proposed a clustering of multi-dimensional data using modified k-means algorithm with artificial neural network. For features extraction exhaustive and heuristic search techniques are used for clustering of data.

S. Moghaddam and M. Ester in [6] described a technique for extract target product that has been commented on in the review. They proposed aspect based

opinion question answering (AQA) system include five following phases; analysis of the question, using a set of part of speech. Expansion of question using the target aspect. Retrieval of high quality of the review based on the target aspects covers.

Finally, it is able to improve the precision and achieve high accuracy of retrieved answers by expanding the question sensing target aspects and retrieving high-quality reviews. Quality filter which are used for filter the high quality of a review is low.

L. Liu et al., in [7] described the explicit and implicit feature technique. The authors proposed a method to be used in extracting the feature and the corresponding opinions, clustering the feature and then orient of an opinion word of the feature.

It is obvious that using the opinion as the feature indicators is ambiguous; this problem can be resolved by relationship between opinion and corresponding feature to improve the precision and recall. The authors also propose a method based on part of speech dictionary to group the opinion. The feature corresponding to the same opinion group can be the candidate set for the implicit feature. There are use small-scale corpus cannot perform well.

X. Yu and Y. Liu in [8] describes a technique for predicting the sales performance of a product by using both sentiments expressed in the public reviews and quality of reviews. They used a probabilistic model known as the sentiment probabilistic latent semantic analysis (S-PLSA).It is used to capture two different factors; 1) revenue of the current day. 2) revenue of the precedence days.

This model further combined to ARSQA model sentiment factor are weighted by the quality have different degree of influence on the prediction PLSA (S-PLSA), S-PLSA model. ARSA, ARSQA. The sentiments expressed in the reviews and the quality of the reviews have a significant impact on the future sales performance of products.

Further improvement can be done on the accuracy using the clustering and classification of reviews based on their sentiments.

III. Methodology

Aspect based product feature categorization including the four steps:

Step-1. Stemming and stop word removing from the reviews sentences.
Step-2. Multi-feature segmentation. Retrieve the aspect and opinion word from the sentences.
Step-3. Determine the Polarity of the opinion of the feature.
Step-4. Cluster the feature of the entity or the product.

Step-1 Stemming and stop word removing

Stemming:

Stemming process is a concern with the mapping and fixing words of the some root stem. This process reduces words by removing suffixes this in turn optimizes the search process and maximizes storage capacity. For example: perfectly - perfect, liked, likes – like.

The advantage or reducing word to a root is to increase the hit rate of identical terms. Stemming is comparatively much simpler heuristic process, which cut's shorts of the words maximizing the storage capacity of a system or a database.

Stop Word Removing:

This process helps in maximizing the storage capacity. In this process the words usually articles, preposition, pronoun etc which may have little lexical meaning with other phases for sentence inside a sentence as, are, the, on, of, with a, about, while, when, who, what, this, that, their, where, who, be, why etc.

Step-2 Multi- Feature Segmentation Method

Different people have different writing style, sometimes people give their reviews of any entity or product, that one sentence of the review contains the multiple aspects, one of the main issues for aspect based opinion polling is to classify the multi-feature based sentences into the single aspect unit. To solve this problem, we use the multi-feature classification (MFC) method. Which classify multi-aspect review sentences into the multi-aspect units? Let S=s1, s2, s3……..sm be a review sentence consisting of m sentences. T=tj, tj+1, tj+2…….tj+n be its sub sentences and A=a1, a2, a3…….an is the aspect of any product or entity and our task is to find the aspect of the given review sentences S.

By calculating the feature (aspect) changes between the sentences, where each segment contains the different feature of the entity or product. For example the feature based classification task. The first review sentences into different sub sentences.

Sentence: Screen is good, backup is high, video player is imagine, connectivity is low.
Segmentation: (screen is good)/ SCREEN-segment‖ (backup is high)/BACKUP- segment‖ (video player is imagine) / video player-segment‖(connectivity is low) CONNECTIVITY-segment.

Our perception to work on the multi feature classification as the different problem of the linear text classification while some linear text classification technique such as frag kou method [16] and dottplotting c99 [17,18], there are some challenges when applying it to multi feature based sentences. First it works well on the document level sentence segmentation rather than sentence level classification whereas our method works on the sentence level.

The main task of this method is to find out the most likely classification with maximum score value T*, two different segment represent the different features (aspect). To determine that the two adjacent represent the same or different aspect we use the T*(S, T) function

$$T^*(S,T)=\begin{cases}1; t_j, t_i+1\in A\\ 0.5; t_i\in A, t_j+1\notin A or\ t_j+1\in A\\ 0; otherwise\end{cases} \quad (1)$$

$$DDF=\frac{1}{C_k}\left[\left(WI+WO+OL\right)\right]\times 100$$

Here, tj represent the sub sentence of the review sentence. There are two adjacent segment express the same aspect whose value is 1. If two given adjacent segment one express the aspect and another not express the aspect whose value is 0.5 otherwise value is 0.

The implementation of multi feature segmentation algorithm is summarized below. Here we use the multi aspect review sentences(S), its sub sentences (T) and collection of different feature (A).

Algorithm-Multi -Feature Segmentation (MFS)

Input- Multi aspect sentences S=s1, s2, s3……sm. Consisting of m review sentences. T=tj, t+1, tj+3…….tj+n be its sub sentences. A=a1, a2, a3…….an, is the collection of different aspect (feature).

FOR each review sentences sm

Where the other aspect occur (words match to the corpus)

Classify the multi feature sentences into the single feature (aspect) unit.

For two adjacent segmentation Calculate the score of sentences

If two adjacent represent same aspect, whose value is 1 and one segment represent aspect and another segment not represent the aspect, value is 0.5 otherwise value is 0.

END FOR
END FOR

Output-Finally, generate multi feature segmentation and the score of the segmentation T*

Feature identification

Feature identification is related to deduce product features out of the tagged text generated by last text. Generally the noun forms of the part of speech give name to the real word. For example, features of product (zoom, battery, image, voice, video etc) these all features are defined by noun. In feature identification process, these words remarkably indicates whole feature of a procedure categorize them into two categories-frequent feature and infrequent feature, feature indication as a process is performed automatically with minimum human intervention.

Relevant feature identification (R1)

People may also use different words to refer to same, the same feature which system with manual annotated will fail to recognize. Moreover people having different opinions may comment on a take of feature of a given product. For example, (picture) and (picture quality) both of them the pair is (picture, picture quality) using the inner word relationship (picture) is common. These features express the same feature of any product or entity. In fact this example shows that the inner word represents relationship between different product feature. We calculate inner word relationship between features using the similarity function mentioned below

$$Sim(f)=2*\frac{p(w1\cap w2)}{p(w1\cup w2)} \quad (2)$$

Where, $p(w1\cap w2)$ represent the probability of common words from given reviews word. $p(w1\cup w2)$ represent the probability of total number of words given in are view sentences.

In other process where features are manually annotated. It is greater advantages that the system will always identify the real feature being frequent or no. The identification of the real feature depends largely on the correctness of previously made annotation. The great disadvantages of this process however is that a large number of annotations have to be made.

Our proposed methodology extracts only noun or noun phrases, adjective, verb from the text. The extracted words are called candidate feature. The set of extracted nouns is matched with a set of sentences containing feature of product in prepared dataset the extract noun forms or to the word net, thesaurus and corpus which ultimately determine the feature of a product.

Irrelevant feature identification:

People have different way to writing the same aspect of The another example, this is costly, as an opinion word costly could be used to describe the aspect (price) of more than one product or object, these opinion words are used to find features which could not be found in the process of frequent feature identification. They have used the phrases, dialogues and the omission words in the process of infrequent feature identification, feature are not directly define. For example the images are absolutely perfect. The application that comes with it is perfect. This is not fit in my pocket. In the above example the first two sentences have one opinion word is common; perfect the second example depends on the first in indirect way, and the first two sentence indicate the feature image and third express the feature size. For this type of review sentences, we use the context information of the given sentences. We use the part-of-dictionary and thesaurus contains the both antonyms and synonyms.

We calculate the weight of each feature by using the point wise mutual information.

$$PMI(f_i w_i)=\log\frac{p(f_j^i w_j^i)}{p(f_j^i)p(w_j^i)} \quad (3)$$

Where $p(f_j^i w_j^i)$ represent the probability of the feature

and the corresponding opinion words. $p(f_j^i)$ is the probability of the feature and $p(w_j^i)$ is probability of the words.

In our approach, we calculate the score of features(direct and indirect) by using the function (1) and the PMI(2).

$$ScoreF(R1,R2) = Sim(f) + PMI(f_i w_i) \quad (4)$$

Step-3 Evaluate the polarity of Opinion at Sentence Level:

Negations rules are follows in determine the sentiment of opinion at the sentence level. A negation words such as no, not and never and some other words that follow pattern such as\stop+ '\ vb- ing" \quit"+\vb-ing can change the orientation of opinion word in the following ways:

1. Negation negative - Positive
2. Negative positive - negative
3. Negative neutral - negative

Some more elaborate example are given below-

- Not perfect - Positive sentiment
- Not good - negative sentiment
- Not work - negative sentiment

The proposed algorithm to evaluate the polarity of opinion at sentence level. Where, we use negation (not, never, nothing, doesn't, don't, haven't, hadn't, can't, shalln't, isn't, willn't etc), OW is opinion word (adjective, adverd, and verb)and opinion polarity (positive, negative and neutral).

Algorithm: polarity of opinion at sentence level.
Input: Review sentence. Here OW represent the opinion word. COUNT (for multiple negation)

- Even number of negation and OW is negative than negative polarity.
- Even number of positive and OW is than negative polarity.
- Even number of negation and neutral than positive polarity.
- Odd number of negation and OW is negative than positive polarity.
- Odd number of positive and OW is negative than negative polarity.
- Odd number of negation and neutral than negative polarity. (for single negation)
- Negation and OW is negative than positive polarity.
- Negation and OW is positive than negative polarity.
- Negation and neutral than negative polarity.
- Positive and neutral than positive polarity.

Output: positive and negative polarity of opinion.

Step-4 Clustering

Clustering is one of the most important research field in data mining in fact clustering is partitioning of data object based methods of data analysis.

The aim of clustering is to minimize intra class similarity and to minimize interclass dissimilarity but this method become complex, when the large dataset are clustered [3].

Many recent studies on the feature categorization aimed to clustering the feature of any product using the semi supervised variant of k-means. The main idea to enhanced the k-mean by using the background knowledge comes from the COP k-mean [15]. Background knowledge means using the constraints in the form of must link and can't link [2, 3] or using the incompatibility constraints [7] to clustering the product but there are three challenges for applying this techniques to categorization, first Mean or Centroid is used to represent the cluster in the k-means method, second it is sensitive to outlier and third large value data object at disrupted the distribution of data.

Our key idea to use the k-medoid clustering method. The k-medoid clustering is most commonly realized through partitioning around medoid (PAM) algorithm. In this method

A medoid is used a central part of data object instead of mean in a cluster medoids based on data object selected randomly represent k-cluster and rest of the data object are fixed in a cluster having medoid similar to that data object.

In our method uses the cannot link constraints to enhance clustering and use the context-dependent information to construct the constraints.

The main function of this algorithm is Manhattan distance by using the Manhattan distance it minimize the overall dissimilarity between the represents of each cluster and its member.

The implementation of proposed method for clustering the product feature categorization consisting of the following steps

The algorithm to identify the features and sentiments, and clustering of given feature in the review.

INPUT: Review Text post by Reviewer on different websites
OUTPUT: Clustering of Product based on positive and negative review.

Step 1. Store review text in array.
Step 2. Remove stop word (i.e. is, am, is, so, already) and stemming (i.e. ing, ed, tional).
Step 3. Break multi feature sentence into single feature unit.
//Statement Store in array
For each Line Review that contains a set of feature's statement
Array [] Line Review
//Words Store in array

Step 4. Break statement into word

For each Line Review that contains a set of feature's statement
Array [] word Line statement

Step 5. Find word from dictionary and feature's
For Line Review in word
If (word ==feature)
Array [index] feature= word;
else if(word ==positive (from Dictionary)
Array [index] positive words= word;
else if word ==Negative (from Dictionary)
Array [index] negative words=w;
end
end

Step 6. Create Data set for store data feature wise
//Store data into dataset based on feature
For each no of count of array
Column1=array [] feature;
Column2=array [] positive;
Column3=array [] negative;
end

Step 7. Apply Clustering Algorithm
Clustering: Group features (i.e. Display, screen)
Define k=2 means create two groups
For each dataset that contains row (feature based)
Dynamically define two features and find minimum cost
Feature is x1, x2, x3, x4, x5
K1=x1, K2=x5 apply for 5 times.
Find minimum cost

Step 8. Clustering: Group 1(x1, x3, x4) Group 2(x2, x5) based

Step 9. Result: Group 1 show that find more positive review and Group 2 show that it feature is get less positive point .

IV. Evaluation and Result Analysis

Dataset

We evaluate various aspect based clustering methods on a wordnet, thesaurus and corpus of computer, camera, mobile, tablet and laptop. All the reviews have been collected from the internet source, and different shopping websites such as www.gmail.com, www.amazon.com. www.jabung.com and www.facebook.com or using the whatsapp.

The dataset contain 4838 reviews for the given product shown in table 1. The dataset contain textual reviews without rating the product and the sentences of textual reviews of different product.

Table 1. Dataset of Reviews

Dataset	Computer	Mobile	Laptop	Tablet	Camera
Revw	1000	760	1050	1000	1028
Sent	1809	1376	2098	3876	1678

In the Table1 the Revw express the total no of review, Sent express total number of sentences in the reviews which are collected from different web sites.

We first calculate the total no. multi feature sentences and remaining single and null sentences. We calculate the different type of sentences from the collection of review sentences given in the dataset.

The dataset contain 10837 sentences which has 3019 multi feature sentences, 1872 null sentences and 5949 single feature sentences shown in Table2. Although multi feature segmentation method is major issue for feature level opinion extraction.

Table 2. Different Type of Sentences

Sent	Computer	Mobile	Laptop	Tablet	Camera
NS	378	177	431	624	262
MFS	456	367	683	965	548
SFS	978	832	984	2287	868

In Table-2, NS represent the null sentences, MFS represent the multi feature sentences, SFS represent the single feature sentences contains in the total number of review sentences.

All the experiments performed on given dataset. The experiments performed for clustering the product feature using our clustering based method. For calculating the performance of the feature categorization or grouping the product using the proposed approach three parameters are used for the evaluation of results.

In the Infrequent and frequent feature identification experiment, we first extract the nouns, adjective, verb and adverb from the review sentences. For divide the review sentences considered the dotting and comma's method studied problem. We considered MFS to extract the multiple features from sentences. In our solution we used the 800 frequently noun, adjective, verb and adverb. They appear frequently in the corpus and found that the learned results are effective.

In the following experiment, threefold cross validation performed and the accuracy, precision and recall three trials for each process shown in table 3, table 4. Only use the linear text segmentation method given the unsatisfactory result on feature based sentence segmentation method. We include the MFC method to obtain the satisfactory result. We have tested three different methods for clustering the product feature categorization are:

Method-1: CA(CW+IW) methods: Here, context word(CW) and inner word (IW) for feature identification, and clustering algorithm (CA) for categorization.

Method-2: CA(CW+OW+IW+POS) methods. We consider context word with corresponding opinion word +inner word relationship + part of speech dictionary + clustering algorithm method. The Context word with corresponding opinion word +inner word relationship + part of speech dictionary for implicit relevant and irrelevant feature recognition and clustering for categorization

Method-3: CA[MFC+(FW+OW)+CW+IW+POS]: We considered MFC+ feature word with corresponding

opinion word+ Context word + inner word relationship + part of speech dictionary + clustering algorithm method adopts. MFC uses for sentence segmentation and feature word with corresponding opinion word+ Context word + inner word relationship + part of speech dictionary for identify the feature and clustering algorithm for categorization of product feature.

We have considered various clustering algorithm (k-mean, k-medoid, cop-kmean, DP), DP is double propagation method. MFS represent the multi feature segmentation method. FW is feature word. OW is opinion word. CW is context word. IW is inner word. POS is part of speech dictionary.

The effectiveness of each product feature categorization method is measured by with precision, recall and accuracy. Table 3, 4 and 5 shows the performance of different clustering the feature categorization approach using the different method.

Table 1 shows performance of different clustering the feature categorization approach using the method-1 as different categorization algorithm with context word and inner words. when we use k-medoid clustering algorithm with background knowledge with context word and inner word for extracting the feature then we grouping highest similarity of feature in one cluster as we can see that the use of k-medoid with background knowledge achieve highest accuracy, precision and recall.

The context word and inner word relationship used for the extracting the feature of the different product feature then clustering the product feature. This method CA (CW+IW) also works best for the other two methods out of three methods.

Table 3. Performance of Different Clustering the Feature Categorization Approach using the Method-1.

	Methods-1	computer	mobile	Laptop	Tablet	Camera
precision	k-mean(CW+IW)	0.47	0.51	0.54	0.45	0.52
	Cop-kmean (CW+IW)	0.51	0.49	0.53	0.52	0.56
	DP(CW+IW)	0.46	0.50	0.52	0.48	0.56
	k-medoid (CW+IW)	0.48	0.51	0.54	0.49	0.55
	OUR(CW+IW)	**0.54**	**0.56**	**0.58**	**0.52**	**0.57**
Recall	k-mean(CW+IW)	0.45	0.57	0.49	0.51	0.50
	Cop-kmean (CW+IW)	0.47	0.49	0.45	0.53	0.50
	DP (CW+IW)	0.44	0.47	0.46	0.48	0.46
	k-medoid(CW+IW)	0.48	0.45	0.53	0.55	0.52
	OUR(CW+IW)	0.56	0.53	0.57	0.56	0.60
Accuracy	k-mean(CW+IW)	0.49	0.47	0.56	0.49	0.57
	Cop-k mean (CW+IW)	0.51	0.50	0.57	0.50	0.59
	DP(CW+IW)	0.42	0.47	0.43	0.46	0.45
	k-medoid(CW+IW)	0.52	0.54	0.58	0.55	0.57
	OUR	0.62	0.66	0.64	0.60	0.61

Table.4 shows the performance of different clustering the feature categorization approach using the method-2. This method achieves better precision, recall and accuracy as compare to method-1. The opinion word and part of speech dictionary performs well for irrelevant feature identification. However, in comparison to method-1, can result in negative effect on irrelevant feature extraction.

Table 4. Performance of Different Clustering the Feature Categorization Approach using the Method-2.

	Methods-2	computer	mobile	Laptop	Tablet	Camera
precision	k-mean (CW+OW+IW+POS)	0.54	0.52	0.58	0.51	0.55
	Cop-kmean (CW+OW+IW+POS)	0.55	0.50	0.59	0.60	0.53
	DPCW+OW+IW+POS)	0.56	0.60	0.54	0.62	0.48
	k-medoid (CW+OW+IW+POS)	0.68	0.54	0.61	0.56	0.55
	OUR (CW+OW+IW+POS)	**0.69**	**0.58**	**0.65**	**0.60**	**0.57**
Recall	k-mean (CW+OW+IW+POS)	0.48	0.57	0.49	0.51	0.53
	Cop-kmean (CW+OW+IW+POS)	0.53	0.59	0.54	0.56	0.59
	DP (CW+OW+IW+POS)	0.46	0.55	0.57	0.55	0.47
	k-medoid (CW+OW+IW+POS)	0.50	0.61	0.66	0.55	0.56
	OUR (CW+OW+IW+POS)	**0.61**	**0.63**	**0.67**	**0.61**	**0.67**
Accuracy	k-mean (CW+OW+IW+POS)	0.51	0.47	0.56	0.49	0.57
	Cop-k mean (CW+OW+IW+POS)	0.53	0.50	0.58	0.51	0.62
	DP (CW+OW+IW+POS)	0.45	0.49	0.50	0.53	0.48
	k-medoid (CW+OW+IW+POS)	0.59	0.66	0.60	0.62	0.64
	OUR	**0.75**	**0.71**	**0.68**	**0.65**	**0.70**

Due to irrelevant feature occur in terms of opinion word and context information of the sentences frequently occurring in the real customer reviews.

Table 5. Performance of Different Clustering The Feature Categorization Approach Using the Method-3.

	Methods-3	computer	mobile	Laptop	Tablet	Camera
precision	k-mean [MFS+(FW+OW)+CW+IW+POS]	0.69	0.56	0.67	0.60	0.68
	Cop-kmean [MFS+(FW+OW)+CW+IW+POS]	0.57	0.52	0.64	0.63	0.57
	DP [MFS+(FW+OW)+CW+IW+POS]	0.58	0.62	0.57	0.69	0.65
	k-medoid [MFS+(FW+OW)+CW+IW+POS]	0.70	0.57	0.68	0.58	0.60
	OUR	0.75	0.69	0.70	0.67	0.78
Recall	k-mean [MFS+(FW+OW)+CW+IW+POS]	0.58	0.60	0.68	0.51	0.53
	Cop-kmean [MFS+(FW+OW)+CW+IW+POS]	0.55	0.58	0.61	0.56	0.59
	DP [MFS+(FW+OW)+CW+IW+POS]	0.51	0.60	0.62	0.55	0.52
	k-medoid [MFS+(FW+OW)+CW+IW+POS]	0.58	0.69	0.71	0.55	0.75
	OUR	0.67	0.77	0.70	0.64	0.78
Accuracy	k-mean [MFS+(FW+OW)+CW+IW+POS]	0.52	0.50	0.66	0.53	0.61
	Cop-k mean [MFS+(FW+OW)+CW+IW+POS]	0.59	0.55	0.67	0.55	0.70
	DP [MFS+(FW+OW)+CW+IW+POS]	0.54	0.58	0.55	0.61	0.52
	k-medoid [MFS+(FW+OW)+CW+IW+POS]	0.68	0.73	0.76	0.68	0.69
	OUR	0.81	0.88	0.77	0.70	0.72

In Table 5., we present the results obtained using method-3. Results obtained shows that, CA [MFC+(FW+OW)+CW+IW+POS] method, outperform than the previous two methods.

We using the multi feature classification (MFC) method outperform on the full stop and dotting method because it is further observed from the customer review sentences that most of the multi feature review sentences generally involve connectivity, battery, application, price etc. feature. We can see that our segmentation techniques contribute to better performance on product feature than state of art approaches. When we include feature word with corresponding to opinion word to measure the inter-relationship between the product feature with the second CA (CW+OW+IW+POS) method obtain the better performance for irrelevant and relevant feature identification than the previous method used for feature identification.

The performance is improved in terms of precision than only considered the context word and inner word to identify the feature word and also the opinion word while adding the part of speech dictionary and multi feature segmentation method, a better result is obtain.

The result shows that the experiment on these MFS+(FW+OW)+CW+IW+POS] technique with the different clustering and categorization method such as k-mean, cop-kmean, DP, k-medoid ,OUR method gives the better results as compare to other method.

Finally we verify the effectiveness of proposed clustering the product feature strategy considering the constraint obtains without any dictionary and automatically, on the performance of feature categorization. We use the constraints as can't link then categorize the product feature from the product review sentences.

All the experimental result that opinion or orientation words are more useful than full context to represent the semantic relationship between the different product features. The group information of opinion word, part of speech dictionary and MFS is very useful in this task especially for irrelevant feature identification and adding k-medoid method with the constraint is also very useful for clustering the product feature.

From the above evaluation and comparison analysis, it can be seen that the proposed methodology provides better result than the other aspect based opinion mining used for the clustering the product feature.

This experiment performed for comparing the performance of clustering based product feature categorization to show that our clustering based product feature categorization gives better or equal than compare to opinion mining based on feature level. This experiment is performed on given dataset. Table 6 shows the performance of clustering based product feature categorization and different feature categorization methods in terms of average precision.

Table6. shows the effectiveness of different methods for feature categorization. This further depicts that our method achieve highest average accuracy 0.7760 percent, by averaging all the five product feature. Therefore, methods that involve MFS and CW+OW+FW with k-medoid (constraint) achieve high average precision, recall and accuracy than those without MFS (FW+OW+IW). The segmentation techniques (MFS) contribute the better performance on computer, tablet, mobile, laptop, camera than their other corresponding techniques. The performance of clustering based product feature categorization system using the k-medoid with constraints algorithm and performance of aspect based opinion mining method for feature categorization achieve in terms of average precision. From above evaluation and comparison analysis, it can be seen that the proposed methodology provides better result than other categorization approach.

Table 6. Evaluation between Various Methods in Terms of Average Precision, Recall and Accuracy.

	methods	(CW+IW)	CW+OW+IW+POS)	[MFS+(FW+OW)+CW+IW+POS]
Average precision	k-mean	0.4980	0.5400	0.64
	Cop-kmean	0.5220	0.5540	0.5860
	DP	0.5040	0.5600	0.6220
	k-medoid	0.5140	0.5880	0.6260
	our	0.5540	0.6180	0.7180
Average recall	k-mean	0.5040	0.5160	0.5800
	Cop-kmean	0.4880	0.5620	0.5780
	DP	0.4620	0.5200	0.5600
	k-medoid	0.5060	0.5760	0.6560
	our	0.5640	0.680	0.7120
Average accuracy	k-mean	0.5160	0.5200	0.5640
	Cop-kmean	0.5340	0.5480	0.6120
	DP	0.4460	0.4900	0.5600
	k-medoid	0.5520	0.6220	0.7080
	our	0.6260	0.6980	0.7760

From the above evaluation, analysis of results that are produced by the given approach (using k-medoid clustering) with constraints for clustering based product feature categorization. In order to calculate the efficiency and effectiveness of the proposed approach, the results of existing approach are compared with the results of product feature categorization using k-medoid algorithm. For calculating the results, three parameters are used. These three parameters are precision, average precision and accuracy. Our proposed approach obtains the better result.

V. Conclusion

In this paper, we have proposed clustering based product feature categorization method in order to produce a relevant summary of any entity.

Opinion word dictionary expensive help to identify the opinion word by providing the antonyms and synonyms of the opinion word, and are also uses the k-medoid clustering method to categorize the product features help to improve the performance of feature classification and categorization method and retrieving rich and high quality of information with increase in precision.

References

[1] S. Momtazi, S. Kazalski, D. Klakow, "A Combined Query Expansion Technique for Retrieving Opinions from Blogs," *Intelligent Systems Design and Applications, ISDA, Ninth International Conference on*, pp. 791-796, 30 Nov 2009.

[2] Y. Choi, Y. Kim, and S. Myaeng, "Domain-specific sentiment analysis using contextual feature generation, "Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, ACM, 2009.

[3] W. J. Jia, S. Zhang, Y.J. Xia, J. Zhang, H. Yu," A Novel Product Features Categorize Method Based on Twice-Clustering," Web Information Systems and Mining (WISM), 2010 International Conference on, vol.1, pp.281,284, 23-24 Oct. 2010.

[4] C. L. Ferm ń, "A knowledge-rich approach to feature-based opinion extraction from product reviews," Proceedings of the 2nd international workshop on Search and mining user-generated contents, ACM, 2010.

[5] J. Zhu, H. Wang, M. Zhu, B.K. Tsou, M. Ma, "Aspect-Based Opinion Polling from Customer Reviews," Affective Computing, IEEE Transactions on, vol.2, no.1, pp. 37-49, 2011.

[6] S. Moghaddam, M. Ester, "AQA: Aspect-based Opinion Question Answering," Data Mining Workshops (ICDMW), IEEE 11th International Conference on, pp.89-96, 11 Dec. 2011.

[7] L. Liu, Z. Lv, H. Wang, "Opinion mining based on feature-level," Image and Signal Processing (CISP),5th International Congress on, pp. 1596,1600, 16-18 Oct. 2012.

[8] X. Yu, Y. Liu, X. Huang A. An, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain," Knowledge and Data Engineering, IEEE Transactions on, vol.24, no.4, pp. 720-734, April 2012

[9] Z. Zhai, B. Liu, H. Xu, & P. Jia, Clustering product features for opinion mining, "Proceedings of the fourth ACM international conference on Web search and data mining", ACM, 2011.

[10] E.Riloff et al "Feature submission for opinion Analysis," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.440-448, 2006.

[11] M.Thomas, B.Pang, L.ee, "Get out the Vote: Determining Support or Opposition from Congressional Floor-Debate Transcripts," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp327-335, 2006.

[12] M.Hu and B.Liu, "mining Opinion Feature in Customer Review," Proceedings of the 9th National Conference on artificial intelligence, 2004.

[13] M.Hu and B.Liu, "Mining and Summarizing Customer Review," Proceedings of the international Conference on Knowledge Discovery and Data mining, pp. 168-177, 2004.

[14] Hu, Mingqin and Bing Liu, "Minning and Summarizing "In proceeding of KDD, 2004.

[15] J.C. Reynar, "An Automatic method of Finding Topic

Boundries," Proceedings of the 32nd annual Meeting Association for CompuationalLinguistics, pp.331-333, 1994.
[16] P. Fragkou, V.Petridis, and A. Kehagias, "A Dynamic Programming Algorithm for Linear text segmentation," proceeding on J.Intelligent Information System, Vol.23, no.2, pp.179-197, 2004.
[17] F.Y.Y.Choi, "Advances in Domain Independent Linear Text segmentation," *Proceeding of the 1st Meeting North* Am.Chapter Association for Computational linguistic, pp.26-33, 2000.
[18] K.W.church, "Char Align P: A Program for Aliging Parallel Texts at the Charcter Level," Proceedings of the 31st Ann.meeting Association for computationallinguisics, pp.1-8, 1993.
[19] C.Scaffidi, K.Bierhoff and et al, Red Opal: product feature scoring from reviews, "proceeding of the 8th ACM conference on Electronic Commerce", pp.182-191, 2007.
[20] A.Popescu and O. Etzioni, "Extarcting Product Feature and Opinions form Reviews," Proceedings of the Conference on Empirical Methods on Natural Language Processing, pp.339-346, 2006.
[21] Kobyashi, Nozomi, K.Inui and Y.Matsumoto, "Extracting Aspect-Evaluation and aspect of Relation in opinion Mining," Proceeding of EMNLP, 2007.
[22] Bahram Izadi, Bahram Ranjbarian, Saeedeh Ketabi, Faria Nassiri-Mofakham, "Performance Analysis of Classification Methods and Alternative Linear Programming Integrated with Fuzzy Delphi Feature Selection", International Journal of Information Technology and Computer Science (IJITCS),Vol. 5, No. 10, September 2013, PP.9-20.
[23] Suneetha Chittineni, Raveendra Babu Bhogapathi, "Determining Contribution of Features in Clustering Multidimensional Data Using Neural Network", IJITCS Vol. 4, No. 10, September 2012, PP.29-36.

Authors' Profiles

Sanjoy Das did his B. E. and M.Tech, Ph.D in Computer Science. Presently, he is working as Assistant Professor, School of Computing Science and Engineering, Galgotias University, India since September 2012. Before joining Galgotias University he has worked as Assistant Professor at Computer Science and Engineering Department, G. B. Pant Engineering College, Uttarakhand, and Assam University, Silchar, from 2001-2008. His current research interest includes Mobile Ad hoc Networks and Vehicular Ad hoc Networks, Distributed Systems, Data Mining.

Bharat Singh, did his MCA from GLA University, Mathura, India and M.Tech in Computer Science from Galgotias University, India in 2011 and 2015 respectively. His current research area includes Data Mining and Opinion mining.

Ms. Saroj Kushwah, did her B.Tech and M.Tech in Computer Science from Uttar Pradesh Technical University, India in 2011 and 2014 respectively. Her current research area includes Data Mining and Opinion mining.

How to cite this paper: Bharat Singh, Saroj Kushwah, Sanjoy Das,"Multi-Feature Segmentation and Cluster based Approach for Product Feature Categorization", International Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.3, pp.33-42, 2016. DOI: 10.5815/ijitcs.2016.03.04