A Two-Level Aspect Categorization For Sentiment Classification Using Association Rule Mining And Deep Hyper Graphs

[1]V.Abinaya , [2]Mrs.K.Krishnakumari
[1]*M.E (CSE) , Avc college of engineering Mayiladuthurai*
[2]*Associate professor, CSE , Avc college of engineering Mayiladuthurai*

Abstract

Sentiment analysis is one of the most active research area in natural language processing and text mining in recent years. It is the computational study of people's opinions, sentiments, attitudes and emotion expressed in written language. Sentiment Classification is a way of analyze the subjective information's the text and then mine the opinions. In the existing sentiment classification system the semantic association among the review is not identified effectively. In this project we are using association rule mining on the context based data from a corpus to separate the aspect categories. A co-occurrence digraph is constructed between aspects and lexicons. Then by using association rule mining proper mapping is performed and polarity can be identified. Finally deep hyper graph model is used to calculate higher order relation among different reviews.

Keywords— *Aspect categorization, Sentiment Analysis, Association Rule Mining, Spreading Activation.*

I. INTRODUCTION

Aspect categorization mean extract all aspect expression of the entities, and categorized these aspect expression into cluster. These aspects can be both explicit and implicit. Aspect-based (or feature-based) sentiment analysis system focus to the detection of all sentiment expression within a given document and the concept and aspect (or features) to which they refer. Sentiment analysis also known as opinion mining which is a natural language processing problem. Natural language processing is related to area of human computer interaction. The task of identifying the opinions of review is called as a sentiment analysis. Opinion may be positive, negative or neutral polarity. Sentiment analysis the process of determining the emotionals tone behind as series of words, used to gain an understanding of the attitudes, opinions, appraisal and emotions expressed by people in written language. Sentiment classification is the special task of text classification whose objective is to classify a text according to the sentiment polarities of opinion it's contain. Sentiment classification performed at different level

- Document level
- Sentence level
- Entity Aspect level

For example: Opinion of a product

"I bought an iPhone a few days ago. It is such a nice phone. The voice quality is clear too. It better than my old Blackberry, which was a terrible phone.

- Documents level sentiment analysis: is this review + or -?
- Sentence-levels sentiment analysis: is each sentence + or -?
- Entity-levels sentiment analysis: is iPhone + or -?.

Document level review classification is too coarse grained and does not provide the desired information. Usually classifying opinion at document level or sentence level is often insufficient for application, because they do not identify opinion of target or aspect. In our project the classification based on aspect level. English is one of the most preferred language to work for natural language processing. This project is based on opinion in English language, does not support other language processing.

This paper contain section1.Literature survey 2: architecture diagram 3: methodology 4: experimental result and its analysis 5: conclusion and future work.

A. Objective

The main aim of the research is concerned with aspect based sentiment analysis, where the goal is to identify the aspect of given target entities using the machine learning techniques. The two level of aspect based sentiment classification are done through association rule mining and deep hyper-graph model. Association rule mining is a procedure to find frequent pattern, correlation, association among the dataset. It aims to predict rules using the co-occurrence term in the review sentence. Deep hyper graph model is based on neural network using supervised machine learning approach. Spreading activation and probabilistic activation method is used to extract aspect terms.

B. *Need of Aspect Categorization in Sentiment Analysis*

- Aspect based Sentiment Analysis refer to the practice of apply Naturals Language Processed ,Text Analysis techniques to identify and extract subjective information from a piece of text.
- Aspect level sentiment analysis gives the detailed and fine grained result when compared to the document and sentence level sentiment analysis.
- Sentiment analysis is one of the active research area in recent years. There is a wealth of informations out there hidden in individual comment, email, tweets, form submission, reviews and the challenge is wrangling of this information and extracting value from it.
- All table should be number with Arabic numeral. Every table should have a captions. Heading should be placed above tables, left justified. Only horizontal lines should be used within a table, to distinguish the column heading from the body of the table, and immediately above and below the table. Tables must be embedded to the text and not supplied separately. Below is an example which the author may find useful.

II. LITERATURE SURVEY

A. *Related Work*

Supervised Learning Based Approach to Aspect Based Sentiment Analysis **Krishna deva et al [1]** - In aspect-based sentiment analysis the aim is to identify the aspects of entities and the sentiment expressed for each aspect. The ultimate goals to generate summaries listing all the aspects and their overall polarity. It give better natural languages understanding and that can help produces more accurate results of SA. The document level review classification is too coarse-grained and does not provide the desired information. **Dimitrios et al [2],** Aspect Based Sentiment Analysis, method can analyze large amounts of unstructured texts and extract coarse- or fine-grained information from review dataset In the text-level ABSA task is fully based on the sentence-level task and the goal was to identify a set of tuples that summarize the opinions expressed in it. These method is well performed for extracting explicit aspect in the review sentence. In that the method only extract the explicit aspect it does not concentrate in the implicit aspect. **Hossein [3],** a new sentiment classification method based on hybrid classification in twitter, the author used a new method for classifying sentiments on Twitter based on combination classification called stochastic gradient descent Hoeffding's tree. The author separate the sentiment analysis in streaming data into two phases that is pre-processing and online process of streaming data. These Prediction of gradient descent has good result when there is more accuracy in average compared with majority of class. **Rigau [4],** Unsupervised Generation of Domain Aspect Terms for Aspect Based Sentiment Analysis, the author proposed simple and unsupervised approach. The goal is to obtains an extended aspect term and opinion word lists. They only extract noun as aspects term and adjective as opinion words, using Stanford parser tools. Double Propagation are used to generate the rule set. These model fails when the category or the aspect terms are more vague or abstract. **Josef Steinberger [5],** Machine Learning Approach to Aspect-Based Sentiment Analysis, consists of four sub-tasks they are aspect term extraction, aspect term polarity, aspect category detection, aspect category polarity. For each subtask author proposed both constrained and unconstrained approach on restaurant review. Conditional random field is used to extract the aspect term in the review. The maximum Entropy classifier model is used for other three task. Finally, the aspect term is extracted and the polarity can be identified using those approach. The performance measure are low for constrained system. **Zheng et al [6]** Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification, the author used unsupervised dependency analysis-based approach to extract Appraisal Expression Patterns (AEPs) from customer reviews. Topic modeling technique is used to explore the association between aspect and sentiment word in the review. It is a sentence-level model so that the accuracy is less. It having many useless AEP's so it leads to the model ineffectiveness. **Zhu et al [7]** Detecting Aspects and Sentiment in Customer Reviews, the author used lexicon based approach towards the restaurant and laptop review dataset. Word cluster is used to reduce data sparsity problem using Brown clustering algorithm. After that the laptop domain is used as a training set to train the SVM classifier using in-domain sentiment lexicon, then the author used restaurant review as a test set for the model. Compare to these two lexicons word aspect association lexicon is to collect all the term having high or moderate association with aspect category. **Wei Wang & Wan et al [8]** Implicit feature identification via hybrid association rule mining, the author used novel hybrid association rule mining to predict many association rules using several complementary algorithms. The best rule in five differents rules set are chosen as the basic rule. the latest rule are used to identify implicit of features. A few more parameters of the hybrid association rule mining so that it is a little hard to control in practical application. **Yu Zhang et al [9]** Extracting Implicits of Feature in Online Customers Review for Opinion

Mining, author used novel co-occurrence association-based method, which aims to extract implicit features in customer review and provide the more kind of comprehensive and fine-grained mining result. These method used Mobile and cloth reviews dataset from Taobao website. These method doesn't work well in cloth domain, because the features of clothes are more complicated and variable. **Chien-Liang Liu et al [10]** Movie-rating and review-summarization system in a mobile environment. The movies-rating information are based to the sentiment classification results. The author collect movie review dataset from Internet Blogs. SVM sentiment classifier, which will classify the reviews into positive or negative classes. The novel based approach on Latent Semantic Analysis (LSA) and PLSA to identify the product features and a statistical approach to identify opinion words. PLSA approach does not work well in product feature identification.

B. Disadvantages of Existing System

- The major challenge in mining aspect and sentiment words without supervision is how to explore the associations between aspect and sentiment words appropriately
- The small amount of annotated data set the result will be inaccurate.
- An early work on Aspect Based Sentiment Analysis (ABSA), the aspects are identified properly which lead to low accuracy compared to human judgment.
- Review sentence having weird grammatical structures, so that the syntactical parser will not be able to extract relevant dependency relation from these sentences.
- The algorithm need sufficient amount of annotating data in order to work properly.
- It doesn't work well when the category or the aspect terms are more vague or abstract.

III. PROPOSED SYSTEM

In this proposed system, both an unsupervised and a supervised method are proposed that are able to find aspect categories based on co-occurrence frequencies. The proposed unsupervised method uses the spreading activation algorithm on the constructed di-graph using the co-occurrence of word in the review sentence. The only required information is the set of aspect categories that is used in the data set and the seed word list for each aspect category. The supervised method on the other hand uses the co-occurrences between words, as well as grammatical relation between the review sentence, and the annotated aspect categories to calculate conditional probabilities from which detection rules are mined. The Stanford Parser is used to extract the grammatical relation between the words. The word in the sentence, such as ''NN'' for a noun or ''JJ'' for an adjective only considered for our work. The position of the corresponding word in the sentence represent the dependency relationship between these two words in the dependency graph. The low frequency word are eliminated for avoid the problem of over fitting. This paper aims to implement an aspect based opinion miner for restaurant domain, which automatically finds important features or aspects (e.g., food, service of restaurants etc.) and its opinion.

IV. SYSTEM ARCHITECTURE

The restaurant review dataset is given as input to the system. The first step is to clean the review dataset using the preprocessing method. In that preprocessing step the stop word, punctuation are removed. After filtering that the set of notion word are obtained. The co-occurrence weighted matrix are constructed using those notion word as a rows and columns and the weight are calculated using the seed word list.

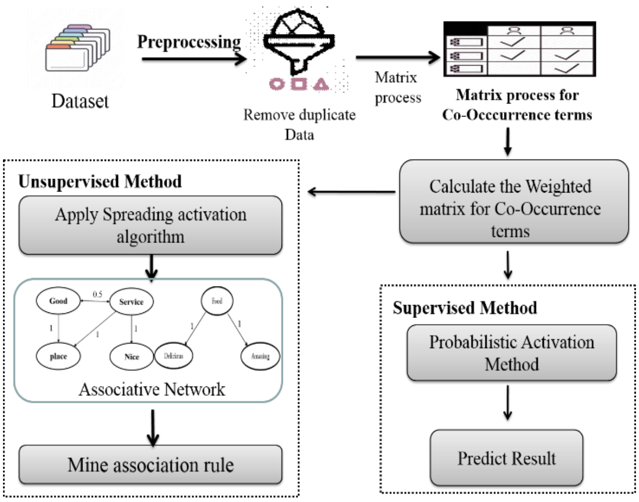


Fig. 1 – System Architecture

In the unsupervised method the spreading activation algorithm are applied to the associative network, which is built using the co-occurrence matrix. The search process of finding an associative network is initiated by giving each vertex an activation value. These initial values determine the area of the search as the activation values are iteratively spread out to other, linked, vertices. After that rules are mined. In the supervised method the Stanford parser tool is used to predict the grammatical relationship between the review sentences Next process is to execute a linear search for optimal thresholds on training data set. The linear search used to find the maximum conditional probability. If the maximum conditional probability is higher than the associated trained threshold the category is assigned to that sentence.

V. METHODOLOGY

There are six main stages:

- Pre-processing Reviews

- Calculating weighted matrix
- Formation of associative network
- Computing Activation Function
- Enumerating Association rules
- Aspect Categorization using probabilistic Activation

Sample Review:
Review 1: The food was delicious.

Review 2: The service is nice.
Review 3: Good service and good place.
Review 4: The food was amazing.

A. *Pre-processing Reviews*

Pre-processing step is required to clean and prepare data for processing. It describes as any types of processing performed on raw data's to prepared it for another processing procedure. Commonly used as a preliminary data mining practice, data pre-processing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. In that the stop words like `the` and `and ` as well as less frequent words are removed from the review dataset because it they add little value for determining the categories in review sentences. After that the list of notion word are obtained from the pre-processing step. Pre-processing is used to identify the notion word occurred in the dataset. The parameter α is used to filter out low occurring lemmas. The low frequency words are removed using the threshold value to avoid the problem of over fitting.
The minimal occurrence threshold

α= 0.005* Number of sentence.

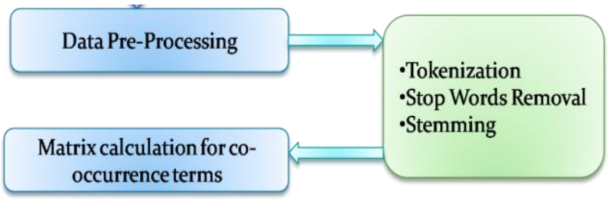


Fig. 2 - Pre-Processing of Review dataset

EXAMPLE:
Sample review is taken from the review dataset and perform the preprocessing step on that data. The result will be like-
Review 1: Food delicious
Review 2: service nice
Review 3: Good service Good place
Review 4: Food amazing

B. *Calculating Weighted Matrix*

After the preprocessing step the list of words are obtained and these words are now considered as a notional word. A co-occurrence matrix is constructed where each entry represents how often the one word Ni that appeared before another word in the same sentence. Each notion word are appeared as the rows and columns in the matrix.

Each entry in this co-occurrence matrix represents the frequency degree of two notional words co-occurring in the same sentence. The co-occurrence digraph G (V, E) is constructed with nodes V and edges E. Each notional word i ∈ N receives its own node i ∈ V. A directed edge (i, j) ∈ E between nodes i and j exists if and only if the co-occurrence frequency X_i is strictly positive. The weight of each edge (i, j) ∈ E is denoted by Wi,j and represents the conditional probability that notional word i co-occurs with notional word j in a sentence after it, given that j is present in that sentence.

The weight is calculated using the below formula-

$$W_{i,j} = (X_{i,j})/N_j$$

Where Xi,j is the co-occurrence frequency of words i and j (word i after word j)and Nj is the frequency of word j.

Table 1 Sample Co-Occurrence Matrix

$X_{i,j}$	Amazing	Delicious	Food	Good	Nice	Place	Service	N
Amazing	0	0	0	0	0	0	0	1
Delicious	0	0	0	0	0	0	0	1
Food	1	1	0	0	0	0	0	2
Good	0	0	0	0	0	0	1	2
Nice	0	0	0	0	0	0	0	1
Place	0	0	0	0	0	0	0	1
Service	0	0	0	1	1	1	0	2

Example:

$$Wservice,good = (X food,amazing)/(N amazing) = 1/1 = 1$$

$$Wservice,good = (X service,good)/(N good) = 1/2 = 0.5$$

$$Wservice,nice = (X service,nice)/(N nice) = 1/2 = 0.5$$

C. Formation of Associative Network

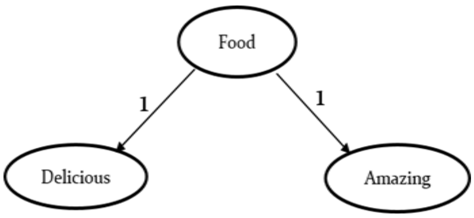


Fig. 3- Associative Network for category food

A network data structure is needed, consisting of vertices connected by links. The vertices are labeled and the links may receive direction and/or weights to model the relations between vertices. The search process of finding an associative network is initiated by giving each vertex an activation value.

Associative network (AN) for category food and service are constructed using co-occurrence matrix and weight between the notion word and aspect term are given below. The associative network is constructed for each category in the restaurant domain using the seed word list of each category. Seed word means that category related word or synonyms of the aspect term.

D. Computing Activation Function

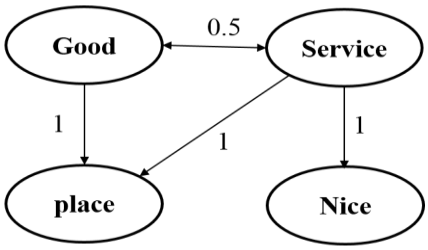


Fig. 4 -Associative Network for category service

To exploit the direct, as well as the indirect relation information between notional words and seed words, the spreading activation algorithm is utilized, which is a method to search for associative networks. Spreading activation has been successfully applied in various fields.

In this project spreading activation is used to find a network of words associated with the category's set of seed words for each category. In the network data structure all notional words receive an initial activation value of zero except for the category's seed words, which receive positive activation values. In the first iterative step of the spreading activation algorithm, these positive activation values are spread out to other words directly related to the seed words, based on the strength of the direct relation.

The formula for the new activation value for one of the vertices j linked to the fired vertex i are given below

$$Ac,j = min\{Ac,j + Ac,i \times Wi,j \times \delta, 1\}$$

Example:

$$Afood,delicious = min\{Afood,delicious + Afood,food \times Wfood,delicious \cdot \delta, 1\}$$

$$= min\{1 + 1 \times 1 \times 0.9\} = min\{1.8, 1\} // closer to 1$$

In the above, $Ac,j > \tau c$, so that the word delicious is in the category food.(c- category)

$$Aservice,place = min\{Aservice,place + Aservice,service \times Wservice,place \cdot \delta, 1\}$$

$$= min\{0 + 1 \times 0.5 \times 0.9\} = min\{0.45, 1\} // closer to 0$$

In the above, $Ac,j > \tau c$, so that the word place is not in the category service. In this way, words that have strong direct relations with the seed words receive high association values.

The following iterative steps will be looking for words with high association values that are then activated and will spread out their activation value to other words directly related to them. In this way, notional words that are indirectly related to one of the seed words are also identified. The end result will be a network of notional words, each with their own activation value, the higher the activation value, the more related the notional word will be to the category.

E. Enumerating Association Rules

Once spreading activation is applied to all categories c ∈ C, matrix Ac,i is obtained, containing, for each notional word i ∈ N, activation values for each category c ∈ C.

From these association values, rules are mined, based on the magnitude of these values. Vertices that have fired are seen as part of the associative network and from each vertex in that network, a rule is mined. Any vertex whose activation value Ac,i is higher than parameter τ c produces a rule [notional word i → category c] that is stored in rule set.

Rule Set:
Review 1: The food was delicious - {food}
Review 2: The service is nice – {Service}
Review 3: Good service and good place - {Service, Place}
Review 4: The food was amazing – {Food}

Since the same word can be present in multiple associative networks, one word might trigger multiple aspect categories. Based on the words in the sentence, a set of rules is triggered and their associated aspect categories are assigned to the sentence.

F. Aspect Categorization Using Probabilistic Activation

The aspect term are categorized using the probabilistic activation method (supervised Learning Method). As a natural language preprocessing step, the dataset are run through the part-of-speech tagger, lemmatizer, and dependency parser of the Stanford CoreNLP

JJ /NN	Food	Service	No. of count
Amazing	1	0	1
Delicious	1	0	1
Good	0	1	1
Nice	0	1	1

The resultant parsed sentence is taken as a input to the probabilistic activation method. Stanford parser is used to identify syntactic relationships between aspect and sentiment words. It is observed that the relationships between aspect and sentiment words are typically represented by the dependency path between them in a dependency graph. In the resultant folder ‘‘NN’’ for category or ‘‘JJ’’ for an notion word is only consider to construct the matrix. After calculating the matrix between the notion word and aspect term. we calculate for each co-occurrence entry Xc,j , with occurrence frequency Yj greater than θ, its associated conditional probability P(c|j), and store it in weight matrix W

$$Wc,j = (Xc,j)/Yj$$

Example:

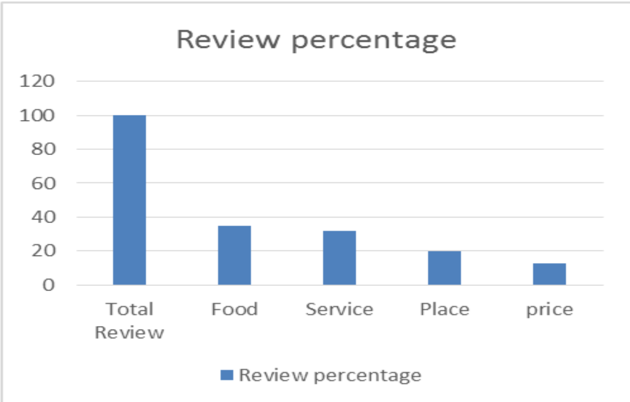
$$WFood,amazing = (X\,food,amazing)/(Y\,amazing) = 1/1 \quad = 1$$

$$Wservice,good = (X\,service,good)/(N\,good) = 1/1 = 1$$

For each category we optimize the thresholds τc value to predict the accurate result. To find theses thresholds a simple linear search is performed, picking the best performing value from a range of values for each different threshold. The selection of one threshold influences the selection of the other three thresholds, all thresholds are optimized together.

VI. EXPERIMENTAL ANALYSIS

For the evaluation of the proposed methods, the training dataset are used. It contains 200 training sentences that are taken from Amazon restaurant reviews. Each sentence has one or more annotated aspect categories. So that each sentence has at least one category and that approximately 5% of the sentences have multiple categories. With 5% of the sentences having multiple categories, a method would benefit from being able to predict multiple categories. This is one of the reasons why association rule mining is useful in this scenario as multiple rules can apply to a single sentence. The restaurant review dataset used in the project has 35% of food related review, 30% of service related review, 20% of place related review and 15% of price related review.



A. Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification models on the dataset for which the true value are known. Calculation of a cross-tabulation of observed and predicted class as associated statistic .they are two type problem, the sensitivity, specificity, positive predictive values and negative predictive as a value is calculated using the positive argument. The confusion matrix for the category food is given below

FOOD n=100	Predicted NO	Predicted YES
Actual NO	True Negative 40	False Positive 20
Actual YES	False Negative 10	True Positive 30

If we were predicting how many food related review are there in total number of training dataset. For example if yes means the food related review are present and no means the review not related to food aspect. Out of 100 the 40 reviews are actually not food related reviews and the prediction value also no and true positive is 30.

The overall accuracy of the two method are calculated using the confusion matrix value. The true positive, true negative and total review are used to calculate the accuracy of the method for aspect categorization.

$$Accuracy = (TP + TN)/Total$$

1) *For Unsupervised Method:*

The accuracy of the unsupervised method has yield 77 % on the dataset set, the method seems to perform well, but the performance strongly depends on the choice of the parameters. In these method, 100 restaurant review sentence are taken to perform the aspect categorization. The four aspect are recognized in the given dataset they are food, place, service and price. In the training set for unsupervised method has 40 food related reviews, 30 place related reviews, 20 service related reviews and 10 price related reviews are there. After performing the activation method on the review dataset the resultant reviews are separated in terms of aspect level. These method gives the result that having 50 review are related to food, 44 reviews are related to place and environment of the restaurant, and the remaining 32 and 7 reviews are related to service and price respectively. But actual value has slight difference compare with predicted value.

2) *For Supervised Method :*

The accuracy of the supervised method has yield 89 % on the dataset set, the method seems to perform well. But the performance strongly depends on the parser. In these method, 100 restaurant review sentence are taken to perform the aspect categorization. The four aspect are recognized in the given dataset they are food, place, service and price.

	Food	Service	Place	Price
Method I	50	32	44	7
Method II	35	40	25	15

In the training set for supervised method has 40 food related reviews, 30 place related reviews, 20 service related reviews and 10 price related reviews are there. After performing the activation method on the review dataset the resultant reviews are separated in terms of aspect level. These method gives the result that having 35 review are related to food, 40 reviews are related to place and environment of the restaurant, and the remaining 25 and 15 reviews are related to service and price respectively. But actual value has slight difference compare with predicted value.

VII.CONCLUSION

In this project two methods are presented for detecting aspect categories that is useful for online review summarization. The first unsupervised method apply the spreading activation on constructed associative network from word co-occurrence data, enabling the use of both direct and indirect relations between words. This results in every word having an activation value for each category that represents how likely it is to imply that category. While other approaches need labeled training data to operate, this method works unsupervised. This method is that a few parameters need to be set beforehand, and especially the category firing thresholds value need to be carefully set to gain a good performance. We have given heuristics on how these parameters can be set. The second, supervised, method uses a rather straightforward co-occurrence method where the co-occurrence frequency between annotated aspect categories and both lemmas and dependencies is used to calculate conditional probabilities. If the maximum conditional probability is higher than the associated trained threshold the category is assigned to that sentence.

VIII. FUTURE WORK

In terms of future work, we would like to investigate how injecting external knowledge would improve the results. While lexicons are a good way of doing that so we have to use the sentiment lexicons. We are especially interested in exploiting more semantic alternatives, like ontologies or other semantic networks. The higher order relationship between the notion word and the aspect term are needed to improve the accuracy of the system. Deep hyper-graph model is basically used to find the high correlation among the sample.The higher order correlation among the notional word and the co-occurrence word are needed to improve the accuracy of the sentiment classification. Deep Hyper graph Model capture the high-level features and reflect the high-order relations among samples. Hyper graph model is used to find correlation among the different reviews. After that comparing those two model to find very higher probability of the model for sentiment classification. The CNN are used to calculate the polarity of the review in an efficient manner.

.

REFERENCES

[1] Schouten, K., van der Weijde, O., Frasincar, F., & Dekker, R. (2018). Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis with Co-occurrence Data. IEEE transactions on cybernetics, 48(4), 1263-1275.

[2] Pannala, N. U., Nawarathna, C. P., Jayakody, J. T. K., Rupasinghe, L., & Krishnadeva, K. (2016, December). Supervised learning based approach to aspect based sentiment analysis. In Computer and Information Technology (CIT), 2016 IEEE International Conference on (pp. 662-666). IEEE.

[3] Zheng, X., Lin, Z., Wang, X., Lin, K. J., & Song, M. (2015). Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. Knowledge-Based Systems, 61, 29-47.

[4] Minab, S. S., Jalali, M., & Moattar, M. H. (2015, November). A new sentiment classification method based on hybrid classification in Twitter. In Technology, Communication and Knowledge (ICTCK), 2015 International Congress on (pp. 295-298). IEEE.

[5] Brychcín, T., Konkol, M., & Steinberger, J. (2014). Uwb: Machine learning approach to aspect-based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 817-822).

[6] Zheng, X., Lin, Z., Wang, X., Lin, K. J., & Song, M. (2014). Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. Knowledge-Based Systems, 61, 29-47.

[7] Kiritchenko, S., Zhu, X., Cherry, C., & Mohammad, S. (2014). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 437-442).

[8] Wang, W., Xu, H., & Wan, W. (2013). Implicit feature identification via hybrid association rule mining. Expert Systems with Applications, 40(9), 3518-3531.

[9] Zhang, Y., & Zhu, W. (2013, May). Extracting implicit features in online customer reviews for opinion mining. In Proceedings of the 22nd International Conference on World Wide Web (pp. 103-104). ACM.

[10] Liu, C. L., Hsaio, W. H., Lee, C. H., Lu, G. C., & Jou, E. (2012). Movie rating and review summarization in mobile environment. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(3), 397-407.