JES Journal of Engineering Sciences

Vol 11, Issue 4 , April/2020
ISSN NO:0377-9254

SUPERVISED AND UNSUPERVISED ASPECT BASED CATEGORY DETECTION USING SENTIMENT ANALYSIS WITH RANDOM FOREST ALGORITHM

Ch. Puspa Satvika[1], P.Tejaswini[2], M.Jyothsna[3] , T.Mahitha[4] , Anusha Papasani [5]
[1,2,3,4] IV B.Tech, Department of Information Technology, Vignan's Nirula Institute of Technology & Science for Women, Peda Palakaluru, Guntur-522009, Andhra Pradesh, India.
[5]Asst.Professor, Department of Information Technology, Vignan's Nirula Institute of Technology & Science for Women, Peda Palakaluru, Guntur-522009, Andhra Pradesh, India.
anoosha.papasani@gmail.com

ABSTRACT

Aspect-oriented sentiment analysis is done in two phases like aspect term identification from review and determining related opinion.. Feature extraction and feature selection techniques contribute to increase the classification accuracy. Feature selection strategies reduce computation time, improve prediction performance, and provides a higher understanding of the information in machine learning and pattern recognition applications etc. This work specifically focuses on aspect extraction from restaurant/hotel review dataset but can also be used for other datasets. In this paper, we propose an unsupervised method to address aspect category detection task without the need for any feature engineering. Our method utilizes clusters of unlabeled reviews and soft cosine similarity measure to accomplish aspect category detection task. This method helps to select relevant features and avoid redundant ones. Initially, the extracted features undergo preprocessing and then the "term-frequency matrix"(TF-IDF) is generated which contains the occurrence count of features with respect to aspect category. In the next phase, different feature selection strategies are applied which includes selecting features based on correlation, weighted term frequency and weighted term frequency with the co-occurence data. this study conducts a sentimental analysis with data sources using with the Random Forest algorithm approach, we will measure the evaluation results of the algorithm we use in this study. The model is good enough. but we suggest trying other algorithms in further research.

Keywords: Aspect-Based Sentiment Analysis (ABSA), Machine Learning (ML), feature selection, Term Frequency-Inverse Document Frequency (TF-IDF), Random forest (RF)algorithm.

1. INTRODUCTION

Due to the quick expansion of the social networking sites, people post their opinions freely[1]. The growth of internet technologies led to increase in online shopping and posting reviews about the products. This helps customers to compare multiple products and gives them further options to choose from. It is a difficult task to analyze products by overall comparison and hence the need to compare products[2]. Comparison can be done on the basis of aspects. ABSA has become a research interest and a challenging task for the researchers. ABSA includes different subtasks namely aspect. term identification, opinion target extraction and corresponding sentiment determination [41]. The sentiment classification is done at three levels like aspect level, sentence level and document level [42].

Following is the example from a restaurant review dataset. Restaurant reviews can have major aspect categories as price, ambiance, food, service, etc. So instead of determining overall review sentiment, it is useful to extract the aspect from review and then determine sentiment for that aspect[3]. In the following example, sentence 1 denotes food aspect category and sentence 2 shows price and food aspect categories [43] [44].

- "The food was great."
- "The food was pricey and not too tasty."

Given a list of pre-defined aspect categories (e.g. 'food' and 'price' in restaurant domain), aspect category detection aims to assign a subset. The aspect

categories may be explicit or implicit. In sentence 2, price aspect is explicit but the food aspect is implicit. The focus of this work is to extract aspect categories from review sentences. For aspect category detection, our method utilizes soft cosine similarity[4].

Firstly, we cluster a set of unlabeled review sentences into k- cluster. Clustering is per- formed based on the Euclidean distance between the average of their word embeddings. Our motivation for using the cluster of sentences is based on the intuition that sentences in the same cluster share similar information about categories they belong to. The similarity between a given sentence and a pre-defined category is defined as the soft cosine similarity between sentence and a set of manually selected seed words corresponding to that category[5]. So, the similarity values can give us information about categories that sentence belongs to. We also define similarity between a cluster and a category as averaging the similarity scores of the sentences in the cluster [46] [47]. Finally, given a test review sentence, scores obtained for the sentence and the nearest cluster to it are interpolated.

These final scores are normalized and used to detect the categories mentioned in the sentence. If the similarity of a category surpasses a threshold, it is as- signed to the sentence. Hence, this is a text categorization problem [48] [49]. This system is trained and tested using Sem-Eval 2014 restaurant review dataset. The reviews in the given dataset had 5 aspect categories like food, ambiance, price, service and miscellaneous[6]. When enough review data is available and aspect categories are defined, then supervised algorithms can be used to forecast the aspect categories. The accuracy of the supervised algorithms(Random Forest) is reliant on the quality of the features extracted and selected [50].

The main challenge when analyzing hotel guest responses is to predict the opinion of an author, expressed in the hotel review, by classifying it as positive or negative feedback[7]. It can be addressed by application of sentiment analysis.

2. LITERATURE SURVEY

Early works for addressing aspect extraction relied on approaches such as identifying frequent nouns and noun phrases using association rule mining, dependency relations, and lexical patterns[8].The SemEval workshops, over the course of three years, has included aspect-based sentiment analysis in their competitions. One of the subtasks introduced dur- ing SemEval is aspect category detection, which our proposed method is going to address[9]. Most of the supervised approaches proposed to address this subtask utilizing machine learning algorithms and train a set of the one-vs-all classifier using hand-crafted features[10].There are only a few unsupervised approaches to address the aspect category detection subtask in the literature. In authors trained a network similar to auto-encoder with attention mechanism to attend to aspect-related words[11][12] In order propose to use a double propagation technique to mine rules based on dependency relations for finding aspect terms of each category[13].Sentiment analysis (SA) is a vastly used term to classify user's opinion using NLP and ML Approaches. Various

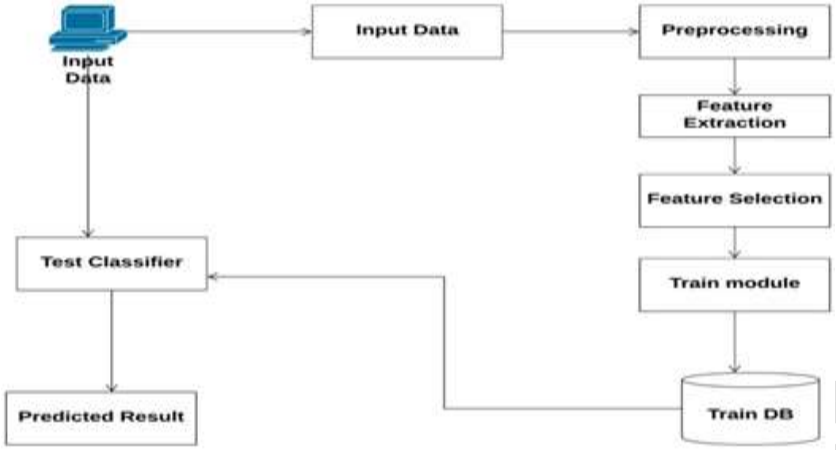

Fig. 1: Supervised learning approach for SA: a basic model view

researchers have used multiple methods for aspect based classification, polarity based classification [14], [15]etc. Product review based sentiment analysis is similar to the proposed sentiment analysis approach. Fig. 1 summaries the basic model of the sentiment classification task.The supervised approach of classification requires a training set and a test set for classification. The classifier is trained on a pre-labeled training set. Then a test set is utilized to verify the model by deriving the class labels of unknown features. Some feature components that are used for feature categorization are unigrams, bigrams NLP based [16] and ontology-based features [17]. Now days, many systems use word dependency-based and ontology-based features [18] to train the classifier. These feature components can be utilized to discover the semantically related words, expressions& sentences.

A variety of machine learning algorithms like Simple Naïve Bayes (NB), Random Forest and Support Vector Machines (SVM)] are used for ABSA of reviews. The NB produces good results if word features are to be considered independent or unrelated to one another. The model depends on Bayesian calculation[19]. Here, lemma features are extracted, and for feature selection different strategies are applied, like feature selection based on term frequency, weighted term frequency, term frequency with correlation, and weighted term frequency with correlation. In this approach, relevant features are selected, redundancy is avoided and unique features are procured. By using these weighted features, a training model is derived. For test sentences, the probability of every aspect class is calculated and the aspect class with the highest probability is the actual aspect label[20][21].

Manual Selection of Category Seed Words

In Existing System K-cluster classification algorithm used on unsupervised data.We manually select a set of 5 seed words for each category (20 in total) to represent the category. Because the anecdotes/miscellaneous category is very unspecific and abstract, we didn't choose any seed words for it[22]For this aspect, following, we would assign this category only to sentences that were not assigned any other category.

Sentence similarity

In order to find the similarity score of a given sentence compares to a category, we utilize soft co- sine measure[23]. For each category, we define the similarity of the given sentence to that category as the average of soft cosine similarity values between the sentence and each of the seed words be- longing to that category. Let x be a given sentence and a_i be the i-th category. We define the $\mathrm{Sim}_{a_i}(x)$ to be the similarity value between a_i and x.

Cluster similarity

A set of unlabeled sentences are acquired from the Yelp dataset challenge [2]. In order to decrease noise samples and since the precision is a more important factor than recall in acquiring true sentences, only sentences that contain at least one of the category

names (eg. 'food', 'service') are selected[24]. Using k-means clustering algorithm, these unlabeled sentences are clustered into k clusters. Clustering is done based on the Euclidean distance between the average of word embedding of sentence words[25], Clustering is unsupervised Learning Algorithm and the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into cluster[26].

3. PROPOSED SYSTEM

Supervised vs. unsupervised learning

Unsupervised learning is when an algorithm is only given input data, without corresponding output values, as a training set[27]. Unlike with supervised learning, there is no correct output values or teachers. Instead, algorithms are able to function freely in order to learn more about the data and present interesting findings. Unsupervised learning is popular in applications of clustering, or the act of uncovering groups within data, and association, or the act of predicting rules that describe the data[28].

Supervised learning models have some advantages over the unsupervised approach, but they also have limitations[29]. The systems are more likely to make judgments that humans can relate to, for example, because humans have provided the basis for decisions. However, in the case of a retrieval-based method, supervised learning systems have trouble dealing with new information[30]. If a system with categories for cars and trucks is presented with a bicycle, for example, it would have to be incorrectly lumped in one category or the other. If the AI system was generative, however, it may not know what the bicycle is but would be able to recognize it as belonging to a separate category. An approach that combines both supervised and unsupervised techniques is called semi-supervised learning[31]. This is when only some of the input data points are labeled with output information. The proposed system as shown in fig 3. processes review dataset and summarize it for better reading. The hotel dataset from the SemEval-2014 Challenge [32] is used in the system. The system consists of three components namely user section, hotel section and admin section. Although, the system is domain specific, it can easily extended to other generalized domains. In proposed system the unlable (unsupervised) data is converted in-to lable (supervised) data. In proposed System Unsupervised methodology have two types of recommendation systems.

a) Content based recommendations

Content-based recommenders treat recommendation as a user-specific classification problem and learn a classifier for the user's likes and dislikes based on product features[33]. In this system, keywords are used to describe the items and a user profile is built to indicate the type of item this user likes.

(1) creating a TF_IDF vectorizer

(2) calculating cosine-similarity

(3) Making a recommendation

b) Collaborative filtering recommendations

The most common technique used for recommendations is collaborative filtering[34]. Recommender systems based on collaborative filtering predict user preferences for products or services by learning pastuser-item relationships from a group of user who share the same preferences and taste.

(i) User-Based Collaborative Filtering

Firstly, we will have to predict the rating that user will give to item[35]. In user-based CF, we will find say k=no of items who are most similar to user . Commonly used similarity measures are cosine, Pearson, Euclidean etc. We will use cosine similarity here which is defined as below:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

(ii) Item-Based Collaborative Filtering

In this approach, similarities between **pair of items** are computed using cosine similarity metric. The rating for target item i for active user a can be predicted by using a simple weighted average as:

$$P_{a,i} = \frac{\sum_{j \in K} r_{a,j} w_{i,j}}{\sum_{j \in K} |w_{i,j}|}$$

where K is the neighborhood of most similar items rated by active user a, and w(i,j) is the similarity between items i and j.Here the item weight is caluculated by TF-IDF.

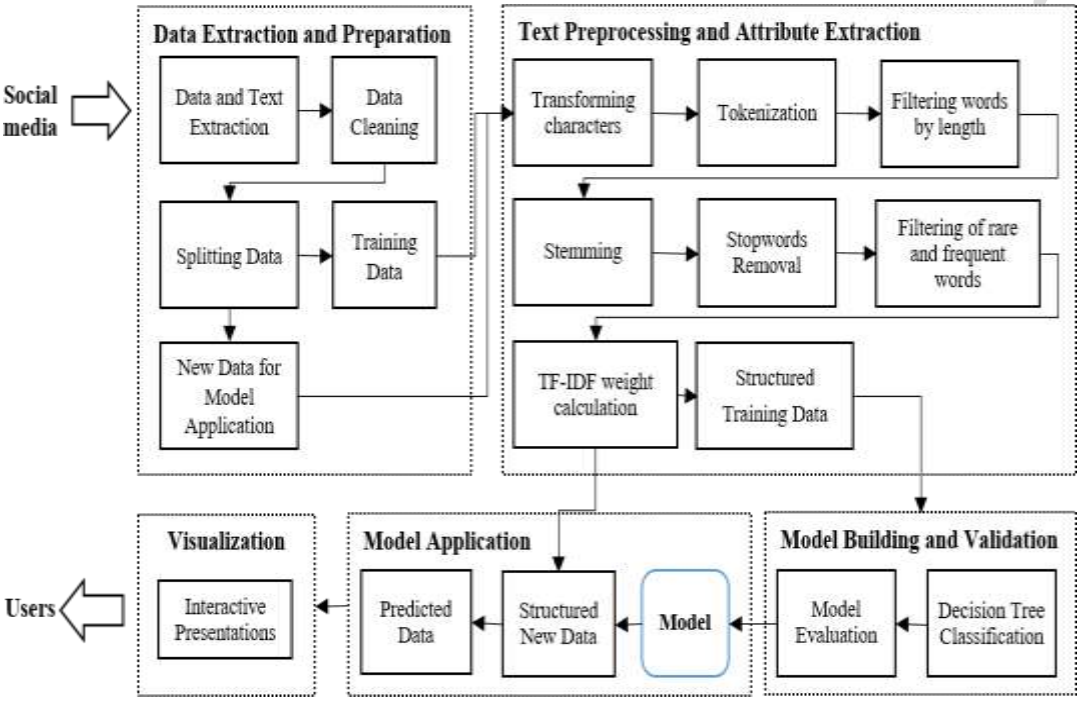

Fig 2: Methodology for Sentiment Classification of Hotel Reviews

Data pre-processing

The review dataset is first pre-processed. The dataset is cleansed so that further processing can be done effectively. All the white spaces and special characters are removed. Also, the reviews are split into each individual sentences. Then, stop words are removed from each sentences as shown in fig.2[36]. This is done by using pre-determined set of stop words. The Porter algorithm is applied for stemming. This results in all the words in their root form and helps in effectively processing them.

Stop word removal

Stop words are regular words that for the most part don't add to the significance of a sentence, in any event for the motivations behind data recovery and normal language handling. These are words, for example, 'the' and 'a[37]. Most web indexes will sift through stop words from hunt questions and records so as to spare space in their file. NLTK accompanies a stop words corpus that contains word records for some languages.

Stemming

Stemming is a procedure to remove appends from a word, winding up with the stem. For instance, the stem of cooking is cook, and a good stemming calculation realizes that the "ing" addition can be removed. Stemming is most usually utilized via web crawlers for ordering words[38]. Rather than putting away all types of a word, a web crawler can store

just the stems, extraordinarily decreasing the measure of file while expanding recovery exactness[39]. A standout amongst the most widely recognized stemming calculations is the Porter stemming calculation by Martin Porter. It is intended to remove and supplant understood additions of English words.

Stop Word Removal

	Processed Review	Stop Words
1	Food was tasty but so expensive.	food tasty expensive
2	To be completely fair, the only redeeming factor was the food, which was above average, but could not make up for all the other deficiencies of Teodora.	completely fair redeeming factor food average not make deficiencies teodora
3	The food is uniformly exceptional, with a very capable kitchen which will proudly whip up whatever you feel like eating, whether its on the menu or not.	food uniformly exceptional capable kitchen will proudly whip whatever feel like eating whether menu not
4	Where Gabriela personaly greets you and recommends you what to eat.	gabriela personaly greets recommends eat
5	For those that go once and do not enjoy it, all I can say is that they just do not get it.	go not enjoy can say just not get
6	Not only was the food outstanding, but the little perks were great.	not food outstanding little perks great
7	It is very overpriced and not very tasty.	overpriced not tasty

Fig 3: Stop Word RemovalLemmatization

Lemmatization is fundamentally the same as stemming, yet is progressively much the same as equivalent word substitution[40]. A lemma is a root word, rather than the root stem. So dissimilar to stemming, you are always left with a substantial word that implies something very similar. Be that as it may, the word you end up with can be completely different.

Algorithm for Preprocessing

Input: a text file containing all the reviews

Output: sentences free from stop words, and stemmed, lemmatized words.

Method: Tokenize the given data set D into sentences S1,S2,S3,…Si again tokenize the sentence into words W1,W2,W3,..Wj

For each word W in a sentence, S look for the presence of it in stop word. If you found it, skip that word else include it and go for another word. Do: FOR EACH SENTENCE in Si DO

Aspect based sentimental analysis

In aspect based sentimental analysis, the main terms to be extracted are the aspects. As it is a complicated process we consider only nouns as the aspect terms. Nouns can be extracted from the chunk grammar.

Data Extraction and Preparation

During data preparation, the text data is preprocessed and unstructured text is turned into a structured format. Firstly, all characters in the example set are transformed to lower case. The text is split into a sequence of tokens, consisting of one single word. Then all tokens which equal a Stop words from the English built-in Stop words list are removed from the text. Stop words are noise words that increase the classification error on new data. The tokens are filtered based on their length, with minimum 3 and maximum 99 characters. Finally, stemming is performed by Porter stemming algorithm applying an iterative, rule-based replacement of word suffixes intending to reduce the length of the words until a minimum length is reached [41].

The tokens are used to generate word vectors numerically representing each example and TF-IDF score of each available word is calculated. The results from preprocessing are in the form of a term document matrix, where each token is now an attribute in a column and each review is an example in a row. The values in the cells are the calculated TF–IDF scores for each word in the word vector creation process. The generated word attributes and their TF– IDF scores are used by the classifier.

Attribute Extraction

The aim of attribute extraction and selection is to select a subset of words occurring in the training set and using only this subset as attributes in text classification. Attribute selection decreases the vocabulary size by eliminating noise or irrelevant words and increases classification accuracy[42]. There are various attribute selection methods based on mutual information, chi-square, Information gain or Gain ratio, frequency-based feature selection. The decision tree algorithm incorporates attribute selection by calculation of Term Frequency-Inverse Document Frequency (TF-IDF) is applied to both, training and testing data sets. It diminishes the weight of terms that occur very frequently in the data set and increases the weight of terms that occur rarely. TF-IDF is calculated through the following formula:

$$TF - IDF = TF_{t,d} * IDF_t$$

where t is a term(attribute) in a document(example) and d is given document(example), where t appears.

Term Frequency (TF) is the ratio between the number of times a term t appears in a given document d (n_t) and the total number of terms in the document (n). The text is split into a sequence of tokens, consisting of one single word.

$$TF - IDF = n * \log_2 N$$

Inverse Document Frequency is the ratio between the total number of documents in the corpus (N_d) and the number of documents that contain the term t (N_t).

The result from applying only TF is that frequent words have higher TF score and infrequent words - lower TF score. TF- IDF takes into account not only the importance of a word in a given document but also its importance in the entire corpus. This technique decreases the weight of frequent words and increases that of rare words in a corpus. The TF-IDF score of a word increases when the number of times the word appears in a document (TF) increases. If the number of documents that contain a word is increased (the word appears more frequently in the corpus), the TF-IDF score of the word decreases, otherwise increases.

TF_IDF Algorithm

Input: sentences free from stop words, and stemmed, lemmatized words.

Output: Aspects that retrieved using tf-idf. Method: Load the input data file.

Calculating the frequency of word in the given input file i.e., number of occurrences of i in j.

TF(w) =Number of times term w appears in a document) / (Total number of terms in the document) Number of documents containing i.

IDF(w) = log_e(Total number of documents / Number of documents with term w in it)

TF-IDF=TF(w)*IDF(w)

Random forest algorithm

Ensemble classification methods are learning algorithms that construct a set of classifiers instead of one classifier, and then classify new data points by taking a vote of their predictions. The most commonly used ensemble classifiers are Bagging, Boosting and Random Forest (RF) [43].

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a

type of learning where you join different types of algorithms or the same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithms of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

RF classifier can be described as the collection of tree- structured classifiers. It is an advanced version of Bagging such that randomness is added to it [15]. Instead of splitting each node using the best split among all variables, RF splits each node using the best among a subset of predictors randomly chosen at that node.

A new training data set is created from the original data set with replacement. Then, a tree is grown using

random feature selection. Grown trees are not pruned [16]. This strategy makes RF unexcelled accuracy [17]. RF is also very fast, it is robust against overfitting, and it is possible to form as many trees as the user wants [18].

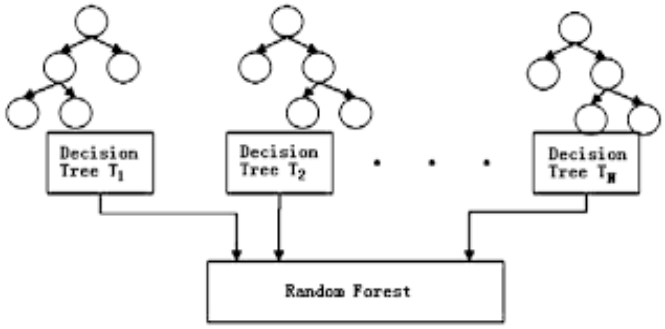

Fig 4: Random Forest –Supervised classification Algorithm

The random forests algorithm (for both classification and regression) is as follows [19]:

1. Draw n_{tree} bootstrap samples from the original data

2. For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when m_{try} = p, the number of predictors.)

3. Predict new data by aggregating the predictions of the e trees (i.e., majority votes for classification, the average for regression).

4. RESULT AND DISCUSSION

The proposed system is implemented in python with windows environment, some inbuilt functions are used during the feature selection as well as extraction. Our proposed aspect based sentiment analysis models reviews were preprocessed to remove unwanted and noisy data. After preprocessing, aspect extraction, identification of interesting aspects, and separation into positive, negative will be performed. Accuracy have been calculated for both the models using Random forest algorithm. The performance metrics indicate that aspect based sentiment analysis model using TFIDF gives higher accuracy on the given input data set. Experimental analysis is explained below

Hotel_Address	Additional_Num	Review_Date	Average_	Hotel_Name	Reviewer	Negative_	Review_	Total_Number	Positive_	Review_	Total_N	Reviewer	Tags	days_sin	lat	lng
s Gravesande	194	8/3/2017	7.7	Hotel Arena	Russia	I am so a	397	1403	Only the	11	7	2.9	['Leisure	0 days	52.361	4.916
s Gravesande	194	8/3/2017	7.7	Hotel Arena	Ireland	No Nega	0	1403	No real c	105	7	7.5	['Leisure	0 days	52.361	4.916
s Gravesande	194	7/31/2017	7.7	Hotel Arena	Australia	Rooms a	42	1403	Location	21	9	7.1	['Leisure	3 days	52.361	4.916
s Gravesande	194	7/31/2017	7.7	Hotel Arena	United K	My room	210	1403	Great loc	26	1	3.8	['Leisure	3 days	52.361	4.916
s Gravesande	194	7/24/2017	7.7	Hotel Arena	New Zea	You Whe	140	1403	Amazing	8	3	6.7	['Leisure	10 days	52.361	4.916
s Gravesande	194	7/24/2017	7.7	Hotel Arena	Poland	Backyar	17	1403	Good re	20	1	6.7	['Leisure	10 days	52.361	4.916
s Gravesande	194	7/17/2017	7.7	Hotel Arena	United K	Cleaner	33	1403	The roor	18	6	4.6	['Leisure	17 days	52.361	4.916
s Gravesande	194	7/17/2017	7.7	Hotel Arena	United K	Apart ho	11	1403	Good loc	19	1	10	['Leisure	17 days	52.361	4.916
s Gravesande	194	7/9/2017	7.7	Hotel Arena	Belgium	Even the	34		No Positi	0	3	6.5	['Leisure	25 days	52.361	4.916
s Gravesande	194	7/8/2017	7.7	Hotel Arena	Norway	The airc	15	1403	The roor	50	1	7.9	['Leisure	26 days	52.361	4.916
s Gravesande	194	7/7/2017	7.7	Hotel Arena	United K	Nothing	5	1403	Rooms v	101	2	10	['Leisure	27 days	52.361	4.916
s Gravesande	194	7/6/2017	7.7	Hotel Arena	France	6 30 AM	75	1403	Style loc	4	12	5.8	['Busine	28 days	52.361	4.916
s Gravesande	194	7/6/2017	7.7	Hotel Arena	United K	The floor	28	1403	Comfy be	6	7	4.6	['Leisure	28 days	52.361	4.916
s Gravesande	194	7/4/2017	7.7	Hotel Arena	Italy	No Nega	0	1403	This hote	59	6	9.2	['Busine	30 days	52.361	4.916
s Gravesande	194	7/4/2017	7.7	Hotel Arena	Canada	The staff	35	1403	It was ve	15	1	8.8	['Leisure	30 days	52.361	4.916
s Gravesande	194	7/3/2017	7.7	Hotel Arena	Italy	No Nega	0	1403	This hote	82	26	10	['Leisure	31 days	52.361	4.916
s Gravesande	194	7/3/2017	7.7	Hotel Arena	United K	Very ste	38	1403	Great or	14	8	6.3	['Leisure	31 days	52.361	4.916
s Gravesande	194	6/30/2017	7.7	Hotel Arena	Ireland	We did n	59	1403	We lover	64	2	7.5	['Leisure	34 days	52.361	4.916
s Gravesande	194	6/29/2017	7.7	Hotel Arena	Netherla	No Nega	0	1403	Public ar	33	4	7.1	['Busine	35 days	52.361	4.916
s Gravesande	194	6/20/2017	7.7	Hotel Arena	Australia	We had i	73	1403	I liked the	48	16	7.5	['Leisure	44 days	52.361	4.916
s Gravesande	194	6/13/2017	7.7	Hotel Arena	United K	Bed was	40	1403	Friendly	17	1	6.3	['Leisure	45 days	52.361	4.916
s Gravesande	194	6/12/2017	7.7	Hotel Arena	France	Our room	92	1403	The bed	25	12	3.8	['Couple	52 days	52.361	4.916

Fig 5: Input Data set

Hotel_Name	Total_Number_of_Reviews	lat	lng	Business	Leisure	Solo	Couple	Group
Hotel Arena	1403	52.360576	4.915968	0	1	0	1	0
K K Hotel George	1831	51.491888	-0.194971	0	1	0	1	0
Apex Temple Court Hotel	2619	51.513734	-0.108751	0	1	0	1	0
The Park Grand London Paddington	4380	51.514218	-0.180903	0	1	0	1	0
Monhotel Lounge SPA	171	48.874348	2.289733	0	1	0	1	0
Kube Hotel Ice Bar	197	48.886570	2.356833	0	1	0	1	0
The Principal London	3150	51.522622	-0.125160	0	1	0	1	0
Park Plaza County Hall London	6117	51.501400	-0.116009	0	1	0	1	0
One Aldwych	259	51.511783	-0.119417	0	1	0	1	0
Splendid Etoile	486	48.874707	2.293676	0	1	0	1	0
Hotel Trianon Rive Gauche	967	48.848768	2.341038	0	1	0	1	0
InterContinental London Park Lane	510	51.503863	-0.150413	0	1	0	1	0

Fig 6: Data set for Content based recommendations

```
: get_recommendations('Milestone Hotel Kensington').head(1

: 469                                Dukes Hotel
  141        London Marriott Hotel Park Lane
  226                        The Pillar Hotel
  672    Mandarin Oriental Hyde Park London
  152                           The Goring
  302                  The Kings Head Hotel
  89                          Marlin Waterloo
  191                           Egerton House
  34                          Hotel Le 10 BIS
  555          H tel Barri re Le Fouquet s
  Name: Hotel_Name, dtype: object
```

Fig 7: output for content based recommendations

```
klean_review1
```

	Hotel_Name	Negative_Review	Positive_Review
0	11 Cadogan Gardens	Thought the prise of drinks at the bar a litt...	We were particularly impressed by the very wa...
1	1K Hotel	Air conditioning in room didn t work and desp...	Location good close to le Marais and 3e arron...
2	25hours Hotel beim MuseumsQuartier	Breakfast not included and buffet really expe...	Cool vintage style in the middle of the museu...
3	41	There wasn t a thing that we didn t like . No...	Its central proximity close to all services a...
4	45 Park Lane Dorchester Collection	More kinds of fruit juice will make the mini ...	Everything here are almost perfect the staffs...
5	88 Studios	Maybe more selection of tea coffee hot chocol...	It was a very nice apartment and the customer...
6	9Hotel Republique	The room was very small but maybe reasonable ...	The bus and metro station are located just in...
7	A La Villa Madame	No Negative, just a better map of surrounding...	The bed was extra comfy the street is really ...
8	ABaC Restaurant Hotel Barcelona GL Monumento	The room looks nice in the pictures with the ...	The room size was bigger than average Barcelo...
9	AC Hotel Barcelona Forum a Marriott Lifestyle ...	no tea and coffee facilities in the room this...	Staff were super friendly and helpful facilie...
10	AC Hotel Diagonal L Illa a Marriott Lifestyle ...	No Negative, No pool coffee machines hard to ...	The rooms were much larger than I had expecte...
11	AC Hotel Irla a Marriott Lifestyle Hotel	The place is ok but room service for cleaning...	Location is ok the staff is helpful, No Posit...
12	AC Hotel Milano a Marriott Lifestyle Hotel	Breakfast could have been included in the roo...	Breakfast was lovely but expensive clean and ...
13	AC Hotel Paris Porte Maillot by Marriott	It was a little off the beaten track Breakfas...	The rooms where excellent everything clean an...
14	AC Hotel Sants a Marriott Lifestyle Hotel	The front desk staff weren t very friendly an...	Right next to the Sants station Rooms are mod...
15	AC Hotel Victoria Suites a Marriott Lifestyle ...	The breakfast was little costly so I did not ...	The rooms were clean and well maintained . P...
16	ADI Doria Grand Hotel	The adress was a little bit misleading A bit ...	The room was very good even the pillows were ...
17	ADI Hotel Polziano Fiera	We could not have an early breakfast on the d...	The public transportation is very close to th...
18	ARCOTEL Kaiserwasser Superior	I booked a suites room and a double room we a...	No Positive, Great location for VIC meetings ...
19	ARCOTEL Wimberger	No Negative, Staff at canteen were not coopar...	Close to public transportation The gentleman ...

Fig 8: Data set for collaborative recommendations

```
get_recommendations('Hotel Saint Petersbourg Opera')

720                                    Hotel Op ra Richepanse
428      H tel Horset Op ra Best Western Premier Collec...
1071                                        Newhotel Roblin
399                                           H tel Bedford
469                                      H tel Westminster
1322                                      The Chess Hotel
164      Best Western Premier Op ra Faubourg Ex Hotel J...
302                                    Edouard 7 Paris Op ra
741                                          Hotel Regina
841                  K K H tel Cayr Saint Germain des Pr s
Name: Hotel_Name, dtype: object

get_recommendations('H tel Westminster')

1226     Saint James Albany Paris Hotel Spa
399                              H tel Bedford
752            Hotel Saint Petersbourg Opera
741                              Hotel Regina
1071                          Newhotel Roblin
800                     Hotel de France Wien
964                  Melia White House Hotel
1473                Washington Mayfair Hotel
1181             Radisson Blu Hotel Amsterdam
1183      Radisson Blu Portman Hotel London
Name: Hotel_Name, dtype: object
```

Fig 9: output for collaborative recommendations

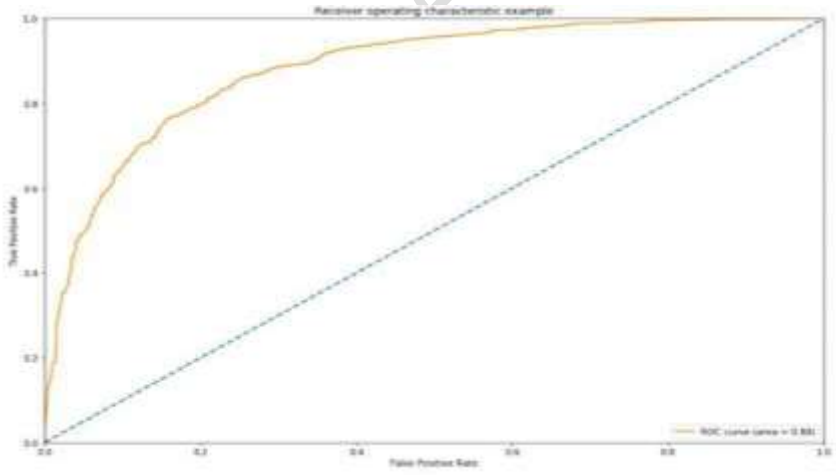

Fig 10: Random Forest classification Evaluation based on ROC curve

5. CONCLUSION

In this work we have presented two methods for detecting aspect categories that is useful for online review summarization. The first, unsupervised, method uses spreading activation over a graph built 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. The major drawback of this method is that a few parameters need to be set beforehand, and especially the category firing thresholds (i.e., co) need to be carefully set to gain a good performance.

We have given heuristics on how these parameters can be set.

REFERENCES

[1]. Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim "Improving Twitter aspect based sentimental analysis using hybrid approach "on 2016.

[2]. Edison Marrese-Taylor, Juan D. Velasquez, Felipe Bravo-Marquez and Yutaka Matsuo "Identifying Customer Preferences about Tourism Products using an Aspect-Based Opinion Mining Approach" 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013.

[3]. Lakshman Narayana Vejendla and A Peda Gopi, (2019)," Avoiding Interoperability and Delay in Healthcare Monitoring System Using Block Chain Technology", Revue d'Intelligence Artificielle , Vol. 33, No. 1, 2019,pp.45-48.

[4]. Gopi, A.P., Jyothi, R.N.S., Narayana, V.L. et al. (2020), "Classification of tweets data based on polarity using improved RBF kernel of SVM" . Int. j. inf. tecnol. (2020). https://doi.org/10.1007/s41870-019-00409-4.

[5]. A Peda Gopi and Lakshman Narayana Vejendla, (2019)," Certified Node Frequency in Social Network Using Parallel Diffusion Methods", Ingénierie des Systèmes d' Information, Vol. 24, No. 1, 2019,pp.113-117.. DOI: 10.18280/isi.240117

[6]. Lakshman Narayana Vejendla and Bharathi C R ,(2018),"Multi-mode Routing Algorithm with Cryptographic Techniques and Reduction of Packet Drop using 2ACK scheme in MANETs", Smart Intelligent Computing and Applications, Vo1.1, pp.649-658. DOI: 10.1007/978-981-13-1921-1_63 DOI: 10.1007/978-981-13-1921-1_63

[7]. Lakshman Narayana Vejendla and Bharathi C R, (2018), "Effective multi-mode routing mechanism with master-slave technique and reduction of packet droppings using 2-ACK scheme in MANETS", Modelling, Measurement and Control A, Vol.91, Issue.2, pp.73-76. DOI: 10.18280/mmc_a.910207

[8]. Lakshman Narayana Vejendla , A Peda Gopi and N.Ashok Kumar,(2018)," Different techniques for hiding the text information using text steganography techniques: A survey", Ingénierie des Systèmes d'Information, Vol.23, Issue.6,pp.115-125.DOI: 10.3166/ISI.23.6.115-125

[9]. A Peda Gopi and Lakshman Narayana Vejendla (2018), "Dynamic load balancing for client server assignment in distributed system using genetic algorithm", Ingénierie des Systèmes d'Information, Vol.23, Issue.6, pp. 87-98. DOI: 10.3166/ISI.23.6.87-98

[10]. Lakshman Narayana Vejendla and Bharathi C R,(2017),"Using customized Active Resource Routing and Tenable Association using Licentious Method Algorithm for secured mobile ad hoc network Management", Advances in Modeling and Analysis B, Vol.60, Issue.1, pp.270-282. DOI: 10.18280/ama_b.600117

[11]. Lakshman Narayana Vejendla and Bharathi C R,(2017),"Identity Based Cryptography for Mobile ad hoc Networks", Journal of Theoretical and Applied Information Technology, Vol.95, Issue.5, pp.1173-1181. EID: 2-s2.0-85015373447

[12]. Lakshman Narayana Vejendla and A Peda Gopi, (2017)," Visual cryptography for gray scale images with enhanced security mechanisms", Traitement du Signal,Vol.35, No.3-4,pp.197-208. DOI: 10.3166/ts.34.197-208

[13]. A Peda Gopi and Lakshman Narayana Vejendla, (2017)," Protected strength approach for image steganography", Traitement du Signal, Vol.35, No.3-4,pp.175-181. DOI: 10.3166/TS.34.175-181

[14]. Lakshman Narayana Vejendla and A Peda Gopi, (2020)," Design and Analysis of CMOS LNA with Extended Bandwidth For RF Applications", Journal of Xi'an University of Architecture & Technology,

Vol. 12, Issue. 3,pp.3759-3765. https://doi.org/10.37896/JXAT12.03/319.

[15]. Chaitanya, K., and S. Venkateswarlu,(2016),"DETECTION OF BLACKHOLE & GREYHOLE ATTACKS IN MANETs BASED ON ACKNOWLEDGEMENT BASED APPROACH." Journal of Theoretical and Applied Information Technology 89.1: 228.

[16]. Patibandla R.S.M.L., Kurra S.S., Mundukur N.B. (2012), "A Study on Scalability of Services and Privacy Issues in Cloud Computing". In: Ramanujam R., Ramaswamy S. (eds) Distributed Computing and Internet Technology. ICDCIT 2012. Lecture Notes in Computer Science, vol 7154. Springer, Berlin, Heidelberg

[17]. Patibandla R.S.M.L., Veeranjaneyulu N. (2018), "Survey on Clustering Algorithms for Unstructured Data". In: Bhateja V., Coello Coello C., Satapathy S., Pattnaik P. (eds) Intelligent Engineering Informatics. Advances in Intelligent Systems and Computing, vol 695. Springer, Singapore

[18]. Patibandla, R.S.M.L., Veeranjaneyulu, N. (2018), "Performance Analysis of Partition and Evolutionary Clustering Methods on Various Cluster Validation Criteria", Arab J Sci Eng ,Vol.43, pp.4379–4390.

[19]. R S M Lakshmi Patibandla, Santhi Sri Kurra and N.Veeranjaneyulu, (2015), "A Study on Real-Time Business Intelligence and Big Data",Information Engineering, Vol.4,pp.1-6.

[20]. K. Santhisri and P.R.S.M. Lakshmi,(2015), " Comparative Study on Various Security Algorithms in Cloud Computing", Recent Trends in Programming Languages ,Vol.2,No.1,pp.1-6.

[21]. K.Santhi Sri and PRSM Lakshmi,(2017), "DDoS Attacks, Detection Parameters and Mitigation in Cloud Environment", IJMTST,Vol.3,No.1,pp.79-82.

[22]. P.R.S.M.Lakshmi,K.Santhi Sri and Dr.N. Veeranjaneyulu,(2017), "A Study on Deployment of Web Applications Require Strong Consistency using Multiple Clouds", IJMTST,Vol.3,No.1,pp.14-17.

[23]. P.R.S.M.Lakshmi,K.Santhi Sri and M.V.Bhujanga Ra0,(2017), "Workload Management through Load Balancing Algorithm in Scalable Cloud", IJASTEMS,Vol.3,No.1,pp.239-242.

[24]. K.Santhi Sri, P.R.S.M.Lakshmi, and M.V.Bhujanga Ra0,(2017), "A Study of Security and Privacy Attacks in Cloud Computing Environment", IJASTEMS,Vol.3,No.1,pp. 235-238.

[25]. R S M Lakshmi Patibandla and N. Veeranjaneyulu, (2018), "Explanatory & Complex Analysis of Structured Data to Enrich Data in Analytical Appliance", International Journal for Modern Trends in Science and Technology, Vol. 04, Special Issue 01, pp. 147-151.

[26]. R S M Lakshmi Patibandla, Santhi Sri Kurra, Ande Prasad and N.Veeranjaneyulu, (2015), "Unstructured Data: Qualitative Analysis", J. of Computation In Biosciences And Engineering, Vol. 2,No.3,pp.1-4.

[27]. R S M Lakshmi Patibandla, Santhi Sri Kurra and H.-J. Kim,(2014), "Electronic resource management using cloud computing for libraries", International Journal of Applied Engineering Research, Vol.9,pp. 18141-18147.

[28]. Ms.R.S.M.Lakshmi Patibandla Dr.Ande Prasad and Mr.Y.R.P.Shankar,(2013), "SECURE ZONE IN CLOUD", International Journal of Advances in Computer Networks and its Security, Vol.3,No.2,pp.153-157.

[29]. Patibandla, R. S. M. Lakshmi et al., (2016), "Significance of Embedded Systems to IoT.", International Journal of Computer Science and Business Informatics, Vol.16,No.2,pp.15-23.

[30]. AnveshiniDumala and S. PallamSetty. (2020),"LANMAR routing protocol to support real-time communications in MANETs using Soft computing technique", 3rd International Conference on Data Engineering and Communication Technology (ICDECT-2019), Springer, Vol. 1079, pp. 231-243.

[31]. AnveshiniDumala and S. PallamSetty. (2019),"Investigating the Impact of Network Size on LANMAR Routing Protocol in a Multi-Hop Ad hoc Network", i-manager's Journal on Wireless Communication Networks (JWCN), Volume 7, No. 4, pp.19-26.

[32]. AnveshiniDumala and S. PallamSetty. (2019),"Performance analysis of LANMAR routing protocol in SANET and MANET", International Journal of Computer Science and Engineering (IJCSE) – Vol. 7,No. 5, pp.1237-1242.

[33]. AnveshiniDumala and S. PallamSetty. (2018), "A Comparative Study of Various Mobility Speeds of Nodes on the Performance of LANMAR in Mobile Ad hoc Network", International Journal of Computer Science and Engineering (IJCSE) – Vol. 6, No. 9, pp. 192-198.

[34]. AnveshiniDumala and S. PallamSetty. (2018),"Investigating the Impact of IEEE 802.11 Power Saving Mode on the Performance of LANMAR Routing Protocol in MANETs", International Journal of Scientific Research in Computer Science and Management Studies (IJSRCSMS) – Vol.7, No. 4.

[35]. AnveshiniDumala and S. PallamSetty. (2016),"Analyzing the steady state behavior of RIP and OSPF routing protocols in the context of link failure and link recovery in Wide Area Network", International Journal of Computer Science Organization Trends (IJCOT) – Vol. 34 No 2, pp.19-22.

[36]. AnveshiniDumala and S. PallamSetty. (2016),"Investigating the Impact of Simulation Time on Convergence Activity & Duration of EIGRP, OSPF Routing Protocols under Link Failure and Link Recovery in WAN Using OPNET Modeler", International Journal of Computer Science Trends and Technology (IJCST) – Vol. 4 No. 5, pp. 38-42.

[37]. VellalacheruvuPavani and I. Ramesh Babu (2019) ,"Three Level Cloud Storage Scheme for Providing Privacy Preserving using Edge Computing",International Journal of Advanced Science and Technology Vol. 28, No. 16, pp. 1929 – 1940.

[38]. VellalacheruvuPavani and I. Ramesh Babu,"A Novel Method to Optimize the Computation Overhead in Cloud Computing by Using Linear Programming" ,International Journal of Research and Analytical Reviews May 2019, Volume 6, Issue 2,PP.820-830..

[39]. Anusha Papasani and Nagaraju Devarakonda,(2016),"Improvement of Aomdv Routing Protocol in Manet and Performance Analysis of Security Attacks", International Journal Of Research in Computer Science & Engineering ,Vol.6,No.5, pp.4674-4685.

[40]. Sk.Reshmi Khadherbhi,K.Suresh Babu , Big Data Search Space Reduction Based On User Perspective Using Map Reduce ,International Journal of Advanced Technology and Innovative Research Volume.07, IssueNo.18, December-2015, Pages: 3642-3647

[41]. B.V.Suresh kumar,Sk.Reshmi Khadherbhi ,BIG-IOT Framework Applications and Challenges: A Survey Volume 7, Issue VII, JULY/2018 pg.no 1257-1264

[42]. P.Sandhya Krishna,Sk.Reshmi Khadherbhi,V.Pavani, Unsupervised or Supervised Feature Finding For Study of Products Sentiment ,International Journal of Advanced Science and Technology, Vol 28 No 16 (2019).

[43]. K.Santhi Sri, Dr.Ande Prasad (2013), "A Review of Cloud Computing and Security Issues at Different Levels in Cloud Computing" , International Journal on Advanced Computer Theory and Engineering Vol. 2,pp 67-73.

[44]. K.Santhi Sri, N.Veeranjaneyulu(2018), "A Novel Key Management Using Elliptic and Diffie-Hellman for Managing users in Cloud Environment", Advances in Modelling and Analysis B,Vol.61,No.2,pp 106-112.

[45]. K.Santhi Sri, N.Veeranjaneyulu(2019), "Decentralized Key Management Using Alternating Multilinear Forms for Cloud Data Sharing with Dynamic Multiprivileged Groups", Mathematical Modelling of

Vol 11, Issue 4 , April/2020
ISSN NO:0377-9254

Engineering Problems,Vol.6,No.4,pp511-518.

[46]. S.Sasikala, P.Sudhakar, "interpolation of CFA color Images with Hybrid image denoising", 2014 Sixth International Conference on Computational Intelligence and Communication Networks, DOI 10.1109/.53 193 DOI 10.1109/CICN.2014.53, pp. 193-197.

[47]. Me. Jakeera Begum and M.Venkata Rao, (2015), "Collaborative Tagging Using CAPTCHA" International Journal of Innovative Technology And Research, Volume No.3, Issue No.5,pp,2436 – 2439.

[48]. L.Jagajeevan Rao, M. Venkata Rao, T.Vijaya Saradhi (2016), "How The Smartcard Makes the Certification Verification Easy" Journal of Theoretical and Applied Information Technology, Vol.83. No.2, pp. 180-186.

[49]. Venkata Rao Maddumala, R. Arunkumar, and S. Arivalagan (2018)"An Empirical Review on Data Feature Selection and Big Data Clustering" Asian Journal of Computer Science and Technology Vol.7 No.S1, pp. 96-100.

[50]. Singamaneni Kranthi Kumar, Pallela Dileep Kumar Reddy, Gajula Ramesh, Venkata Rao Maddumala, (2019), "Image Transformation Technique Using Steganography Methods Using LWT Technique" ,Traitement du Signalvol 36, No 3, pp. 233-237.