Tourism Review Sentiment Analysis using Lexicon Features and Machine Learning Approach

Priya Rai[1] & Ramratan Ahirwal[2]
[1]M.tech Scholar, Department of Computer Science & Engineering, SATI,Vidisha, India[1]
[2]Assistant Professor, Department of Computer Science & Engineering, SATI, Vidisha, India[2]

Received: August 22, 2018 Accepted: October 13, 2018

ABSTRACT *In recent years for text sentiment analysis lexicon-based approaches has been used significantly. Many research work have been used lexicon method for review sentiment analysis. In this paper, a new tourism review sentiment lexicon generation method is proposed using Part-Of-Speech (POS) tags in the lexicon. Training review data is selected from an unlabeled corpus based on their word scores as assessed by the SentiWordNet (SWN). Subsequently, a sentiment analysis framework is proposed based on a complete lexicon that uses the machine learning approach to analyze the review class, both positive and negative. For simulation analysis a tourism review dataset is prepared by collecting different reviews from different websites. The proposed algorithm is analyzed using three classifiers i.e. support vector machine (SVM), k nearest neighbor (KNN) and random forest (RF). After simulation results KNN gives about 80% accuracy, SVM gives about 89% accuracy and RF gives 93% accuracy. The results show that the sentiment analysis framework based on the proposed lexicon is effective for tourism records that use a random forest classifier.*

Keywords: *Lexicon-based approach, Sentiment analysis, SentiWordNet, Machine Learning, Accuracy.*

I. Introduction

Emotions are an important aspect of interaction and communication between individuals. The exchange of emotions through text messages and contributions from personal blogs presents the informal nature of the challenge posed by writing for research. The extraction of the emotions of the text is used to determine the human-machine interaction that controls communication and many others [1] - [3]. Emotions are expressed also through language, the primary emotions of a person's face and text. Emotions are also expressed by a word or by many words.

Sentence level emotion detection technique plays a vital role to trace emotions or to look out the cues for generating such emotions. Sentences are the essential info units of any document. For that reason, the document level feeling detection technique depends on the feeling expressed by the individual sentences of that document that in turn depends on the emotions expressed by the individual words.

The present era of Internet has become a huge Cyber Database which hosts gigantic amount of data which is created and consumed by the users. The database has been growing at an exponential rate giving rise to a new industry filled with it, in which users express their opinions across channels such as Facebook, Twitter, Rotten Tomatoes and Foursquare. Opinions which are being expressed in the form of reviews provide an opportunity for new explorations to find collective likes and dislikes of cyber community.

In recent years, along with the internet popularization and the dynamic development of ecommerce, more and more people are shopping online. There are a large number of users comment on what they purchase online each day. The reviews provide important guidance for both customers and businesses. Customers make their decisions based on existing reviews, and business companies discover the problems of their products or services from the attitudes of users in the reviews [4]. Sentiment analysis is the computational study of people's opinions, sentiments, emotions, and attitudes [5].

Opinion mining or sentiment analysis can be described as the process of automatically extracting and analyzing the opinions, sentiments, thoughts and feelings of opinion writers on a specific target. This target could be a product, some issue like politics, economics, events, phenomena, services, etc.) which defines sentiment to be a personal positive or negative feeling.

Sentiment analysis methods can be generally divided into two categories, machine learning and lexicon-based methods. The former uses machine learning techniques for sentiment polarity classification. These kinds of methods usually need a lot of labeled training data. However, collecting sufficient labeled data is a challenge in itself. Lexicon based methods utilize sentiment lexicons to compute sentiment scores of given reviews. Then they group the scored reviews into positive or negative categories by the sentiment scores.

Many researcher illustrated that lexicon based method is the method for constructing sentiment lexicon. Its generation is divided into two main steps i.e. dictionary-based and corpus-based approaches. In [6] author developed domain-specific sentiment lexicon which results better. For example, long is a positive word when it is used to describe the phone's standby time in a review that comments on a phone, but it is a negative word when it is used to describe the time that a printer prints a paper. Despite a significant amount of research, there is still no an effective method for discovering and determining domain-dependent sentiment lexicon [7].

Sentiment analysis (SA) is the process of identifying and classifying users' opinions from a piece of text into different sentiments—for example, positive, negative, or neutral—or emotions such as happy, sad, angry, or disgusted to determine the user's attitude toward a particular subject or entity. Sentiment analysis (SA) plays an important role in many fields including tourism, where tourists feedback is essential to assess the tourist places all over the world [8]-[10]. Many tourist websites obtain such feedback via a tourist response system for overall journey of them which helps other tourists to decide where to go for refreshment with family or friends. So, such reviews gives effective affords while deciding tourist places. So, opinion mining or sentiment analysis tools helps in this field.

Sentiment Analysis is a technology that will be very important in the next few years. With opinion mining, we can distinguish poor content from high quality content. With the technologies available we can know if a movie has more good opinions than bad opinions and find the reasons why those opinions are positive or negative. Much of the early research in this field was centered around product reviews, such as reviews on different products on Amazon.com [1], defining sentiments as positive, negative, or neutral. Most sentiment analysis studies are now focused on social media sources such as IMDB, Twitter [2] and Facebook, requiring the approaches be tailored to serve the rising demand of opinions in the form of text. In this paper, we follow a lexical approach [3] using the SentiWordNet [4] to determine the overall polarity of the tourism review. We analyze and study the features that affect the sentiment score of the tourism review text. Also, we use the state of the art classification algorithms for the evaluation of performance and accuracy of the approach used. With opinion mining, we can distinguish poor content from high quality content based on the methodology employed sentiment classification can be either learning based or lexicon based. Learning based approaches view opinion mining as a classification task performed using machine learning algorithms.At the same time lexicon based approaches use predefined lexicons or dictionaries.Major steps involved in sentiment analysis include data gathering, preprocessing, aspect identification, feature extraction and sentiment classification. An opinion mining process can be done at different levels like document level, sentence level, phrase level, tweet level or aspect level. Complexity of machine learning based sentiment analysis system depends upon the number and nature of features selected for the design. Many rresearcher proposed an efficient machine learning based sentiment analysis system with reduced feature set by identifying key terms and relevant lexicons.

The existing methods have some deficiencies:

1. These methods only apply to some specific domains in which emotions are used frequently
2. They need human-annotated data
3. The generated sentiment lexicon contains more positive or negative words.

In this paper, a domain-specific sentiment is proposed lexicon generation method and a sentiment analysis framework based on the generated domain-specific sentiment lexicon.

II. Related Work

English is the most popular language for research in Natural Language Processing. Most approaches used in this area are:

- Subjective Lexicon
- Machine Learning

A. Subjective Lexicon Approach

Lexicon approach depends on finding opinion lexicon which analyzes sentiment of text. This approach has 2 methods:- Dictionary based and Corpus based. There are 3 main approaches in finding opinion list. Manual approach is very time consuming so it is combined with either of these two.

Hindi language is scarce due to limited resources till now.

There are three popular methods for generation of subjective lexicon:-

- Use of Bi-lingual dictionary[2]
- Machine Translation[2]
- Use of Wordnet[4]

B. Machine Learning Approach

In such way total feature vector is generated for each review using features. These features are further classified by using classifiers. For each

extracted features of review emotion classification algorithm is applied on different set of inputs. Different classifiers are such as SVM, Neural Network, KNN, Random Forest etc.

Some of the contribution in this field are discussed below:

Sruthi S et al. [1] propose an entity recognition method in the preprocessing stage to eliminate the irrelevant information from the reviews. Performance of SVM is about 95% and proposed sentiment analysis system efficient in terms of time and cost. More features are required based on context.

Tirath et al. [2] extracted features which are strongly effective in deciding the extremity of the movie reviews and used computation linguistic methods preprocessing of the information. Six classification techniques are analyzed on this technique. Found that Random Forest outperforms an accuracy of 88.95%. NLP based feature extraction is not properly discussed.

Md Shad Akhtar et al. [3] proposed a deep learning framework for review sentiment analysis. In this work features are selected using multi-objective optimization (MOO) framework. These optimal features are classified using SVM on four Hindi datasets covering varying domains.

K. M. Anil Kumar et al [4] proposed a machine learning approach for sentiment analysis from Kannada documents. For conversion of English language into kannada language translation tool was applied. Further part of speech tagger is used to implement adjective analysis. The polarity is considered as the difference between the positive and negative counts.

Namita Mittal et al [5] developed an efficient approach based on negation and discourse relation for predicting sentiment. They improved HSWN by adding more opinion words to it. They proposed rules for handling negation and discourse that affected the prediction of sentiments. 80% accuracy was achieved by their proposed algorithm.

M. Farhadloo et. al. [6] proposed multiclass sentiment analysis for English language using clustering and score representation. The model used aspect level sentiment analysis. Bag of nouns was preferred instead of bag of words to enhance clustering results, score representation and more accurate sentiment identification.

V.K. Singh et al. [7] performed experimental analysis on SentiWordNet approach for performance evaluation for document level sentiment arrangement of Movie audits and Blog posts using Naive Bayes and SVM classifier. Achieved approx. 80% accuracy. Restriction with this aspect-level implementation is that it is domain specific.

Das and Bandopadhya[8] gave four strategies to predict sentiment of word. First strategy proposed by them was an interactive game which returned annotated words with their polarity. In second strategy, they use bi-lingual English and other Indian Language dictionaries to predict the polarity. In third approach, they use wordnet and synonym-antonym relation to predict the polarity. In fourth approach, polarity is determined by learning from pre-annotated corpora.

Joshi et al. [10] proposed fall back strategy for Hindi Language. Their strategy follows three approaches: In-Language Sentiment Analysis, Machine Translation, Resource based sentiment analysis. They developed Hindi SentiWordnet(HSWN) by replacing words of English SentiWordnet by their Hindi Equivalents. Final accuracy achieved by them is 78.14.

Michelle Annett et al. [11] proposed an approach based on Support Vector Machines for sentimental analysis on movies reviews and compared with machine learning (ML) approaches. Achieved 60% accuracy and concluded that lexical approach for sentiment analysis is quite efficient. Achieved very low performance accuracy.

The ultimate goal of sentiment analysis can be generally summarized as identifying sentiment or opinion labels of given texts. Depending on the types of final label, problems are usually divided into sentiment classification and emotion/subjectivity identification.

III. Proposed methodology

In this section, the proposed method is illustrated. First, tourism review dataset is prepared which is further used for sentiment analysis. The proposed algorithm mainly consists of three phases i.e. Dataset Preparation, Feature Extraction and Classification which are elaborated below:

A. *Dataset Preparation*

The more disciplined you are in handling of data, the more consistent and better results you can achieve. The process for getting data ready for a machine learning algorithm can be summarized in two steps:

Step 1: Data Acquisition

Step 2: Text Preprocessing

1) Data Acquisition

The data preparation method has been shown in Figure 1. That represents the data acquisition for text review nalysis.

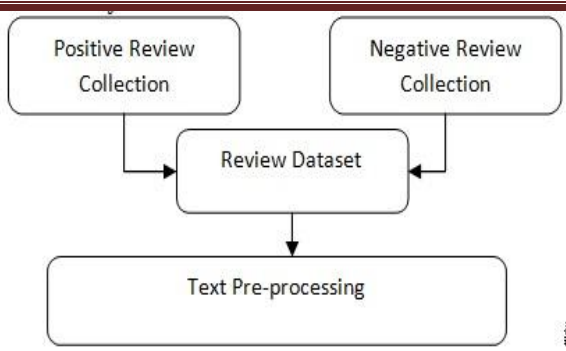

Figure 1: Data Acquisition for Proposed Work

Data Acquisition is performed by collecting reviews of different persons from different location of Trip Advisor. Total 450 reviews have been collected for positive and negative reviews.

2) Text Preprocessing

Text preprocess is mandatory for sentiment analysis in order to deal with texts published by users on social media platforms. The text preprocessing in the proposed method consists of following parts as follows:

1. Remove URLs and special symbols URLs in user reviews do not convey any emotional tendencies.
2. Some special symbols (such as () @ # $%, etc) are often used in the reviews, while these symbols do nothing but interfere with the sentiment classification. In order to avoid these disturbances, special symbols are removed.

B. *Feature Extraction*

Transformation of input data into a set of features. Features are distinctive properties of input patterns that are usable for machine learning approach for classification. In this paper following features are extracted (shown in figure 2):

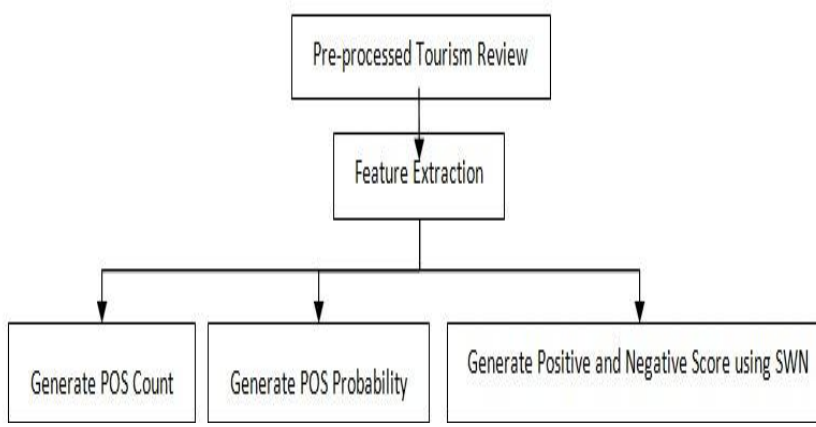

Figure 2: Feature Extraction for Proposed Work

i. Generation of POS count

It is a type of feature extraction in which count of part of speech is extracted from given text reviews. For this Part-of-Speech Tagger (POS Tagger) is used to assign part of speech to each word in the text (and other tokens), such as noun, verb, adjective, etc. It is very useful to the review sentiment analysis. As this process makes use of Penn POS Tags and assigned reviews into 11 parts of speech which is shown in table 1.

Table 1: Look-up Table for POS

Penn POS Tag	Description	Equivalent Proposed Algorithm Tag
JJ	Adjective	ADJ
JJR	Adjective, Comparative	ADJ
JJS	Adjective, Superlative	ADJ
NN	Noun, Singular	NN
NNS	Noun, Plural	NN
NNP	Proper Noun, Singular	NNP
NNPS	Proper Noun, Plural	NNP
RB	Adverb	ADV
RBR	Adverb, Comparative	ADV
RBS	Adverb, Superlative	ADV
VB	Verb	VB
VBD	Verb, past tense	VB
VBG	Verb, gerund or present participle	VB
VBN	Verb, past participle	VB
VBP	Verb, non-3rd person singular present	VB
VBZ	Verb, 3rd person singular present	VB
DT	Determiner/ Article	DT
IN	Preposition	PP
PR	Pronoun	PN
CC	Coordinating conjunction	CC

UH	Interjection	INJ
FW, MD, TO, PDT	Foreign word, Modal, to, Predeterminer	OTH

Let's take an example for demonstration of part of speech tagging. For example, "Close to the bus stand and adequate food options". After using Penn POS Tags, the given example has following POS tagging:

Close	Verb
To	Other
The	Determiner
Bus	Noun
Stand	Noun
And	Conjunction
Adequate	Adjective
Food	Noun
Options	Noun

ii. Generation of POS probability of Occurrence

After POS tagging count of each part of speech is determined and their probability is determined as a feature vector. For POS count in above example is as follows:

Verb	1
Other	1
Determiner	1
Noun	4
Conjunction	1
Adjective	1
Proper Noun	0
Adverb	0
Pronoun	0
Interjection	0
Preposition	0

For probability count of the POS is determined as:

$$\text{Probablity\_POS} = \frac{\text{Count of POS}}{\text{Total Number of words in review sentence}} \quad \text{(i)}$$

For POS probability in above example is as follows:

Verb	1/9 = 0.11
Other	1/9 = 0.11
Determiner	1/9 = 0.11
Noun	4/9 = 0.44
Conjunction	1/9 = 0.11
Adjective	1/9 = 0.11
Proper Noun	0/9 = 0
Adverb	0/9 = 0
Pronoun	0/9 = 0
Interjection	0/9 = 0
Preposition	0/9 = 0

iii. Generation of Positive and Negative Score using SWN

SentiwordNet is used to calculate the total positive and negative score of the sentence or review. By using SentiwordNet dictionary in this research work positive score and negative score of each word in sentence is calculated and after adding positive and negative score of all words, overall positive and negative score of the sentence is calculated. The calculation of positive and negative score of the sentence is determined as follows:

$$\text{Positive\_Score\_Sentence} = \frac{\text{Sum of Positive Score of Each Word}}{\text{Total Number of words in review sentence}} \quad \text{(ii)}$$

$$\text{Negative\_Score\_Sentence} = \frac{\text{Sum of Negative Score of Each Word}}{\text{Total Number of words in review sentence}} \quad \text{(iii)}$$

For Generation of Positive and Negative Score using SWN in above example is as follows:

Word	Pos Score	Neg Score
Close	0.65	0.58
To	0.2	0.2
The	0.1	0.1
Bus	0.45	0.45
Stand	0.5	0.5
And	0.1	0.1
Adequate	0.98	0.54
Food	0.4	0.45
Options	0.54	0.43

So, the overall positive score of the sentence is 0.435 and negative score is 0.372.

After finding POS count, POS probability, positive score and negative score, the feature vector is formed whose sample is shown in below table 2.

Table 2: Sample Feature Vector

C_DT	C_NN	C_PP	C_PN	C_ADJ	C_Adv	C_Vb	C_PNN	C_CC	C_Int	C_Oth	Prob_DT	Prob_NN	Prob_PP	Prob_PN	Prob_ADJ	Prob_Adv	Prob_Vb	Prob_PNN	Prob_CC	Prob_Int	Prob_Oth	Pos_Score	Neg_Score	Class
0	2	0	1	1	0	0	3	1	0	0	0.00	0.25	0.00	0.13	0.13	0.00	0.00	0.38	0.13	0	0	1.3252	1.3904	1
0	2	0	0	0	0	0	0	1	0	0	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0	0	1.3237	1.3893	1
1	3	1	1	1	1	2	0	1	0	0	0.09	0.27	0.09	0.09	0.09	0.09	0.18	0.00	0.09	0	0	1.9874	1.7662	1
1	1	2	1	2	1	1	1	1	0	0	0.09	0.09	0.18	0.09	0.18	0.09	0.09	0.09	0.09	0	0	2.2347	2.041	1
1	2	0	2	2	1	4	0	2	0	0	0.07	0.14	0.00	0.14	0.14	0.07	0.29	0.00	0.14	0	0	1.3246	1.495	1
3	6	2	2	2	1	3	1	1	0	0	0.14	0.29	0.10	0.10	0.10	0.05	0.14	0.05	0.05	0	0	0.45175	1.0816	0
1	6	3	0	2	2	5	1	0	0	0	0.05	0.30	0.15	0.00	0.10	0.10	0.25	0.05	0.00	0	0	0.24799	0.49236	0
1	6	1	1	2	2	2	0	1	0	1	0.06	0.35	0.06	0.06	0.12	0.12	0.12	0.00	0.06	0	0.06	0.30007	0.86894	0
1	3	0	0	3	0	3	1	1	0	1	0.08	0.23	0.00	0.00	0.23	0.00	0.23	0.08	0.08	0	0.08	0.50017	0.54626	0
1	3	1	0	1	1	1	2	1	0	0	0.09	0.27	0.09	0.00	0.09	0.09	0.09	0.18	0.09	0	0.00	0.4984	0.78858	0
5	9	4	1	2	3	5	1	1	0	3	0.15	0.26	0.12	0.03	0.06	0.09	0.15	0.03	0.03	0	0.09	0.17565	0.48061	0
0	1	0	0	3	1	1	2	2	0	0	0	0.1	0	0	0.3	0.1	0.1	0.2	0.2	0	0	0.389	0.393	0

C. *Classification*

Data classification is the process of sorting and categorizing data into various types, forms or any other distinct class. Data classification enables the separation and classification of data according to data set requirements for various objectives.

In this research work, after feature extraction the feature vector is divided into training and testing ratio and sent to classifiers such as Support Vector Machine (SVM), Random Forest (RF) and k-NN (k-nearest neighbour). SVM, k-NN and RF are supervised learning classifier. SVM classifies data into different classes by identifying a hyperplan (line) that separates learning data into classes. The hyperplane's identification, which maximizes the distance between classes, increases the probability of generalizing secret data.

k-NN is one of the simplest classification algorithms. Determine the parameter k which is number of nearest neighbours. When there is new data point to classify, then its k nearest neighbours is find out from the training data.

Random forest is learning method for classification. Multiple decision trees are constructed at training time and outputting the classes or prediction. Random forest applies bootstrap aggregation technique which decorrelates the trees by showing them different training sets.

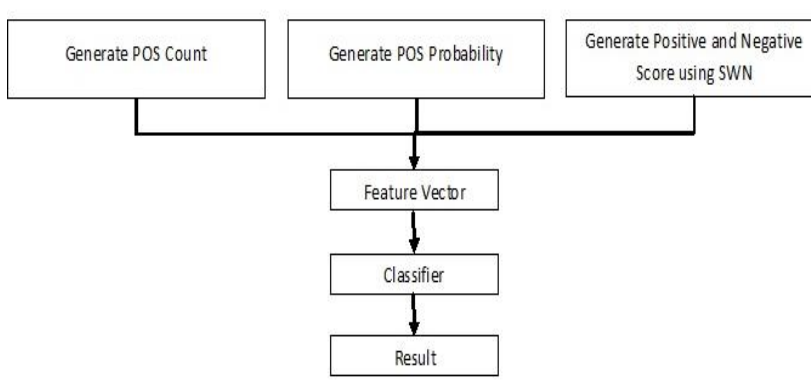

Figure 3: Feature Classification for Proposed Work

Overall proposed algorithm is shown in figure 4 which include three phases i.e. Dataset Preparation, Feature Extraction and Classification.

IV. Result Analysis

In this paper, the tourism review dataset is prepared and simulation is performed using proposed algorithm. For performance evaluation, SVM, KNN and RF classifier are used which are discussed below:

Support Vector Machine

This is for classification and regression problems. SVM classifies data into different classes by identifying a hyperplane (line) that separates learning data into classes. The hyperplane's identification, which maximizes the distance between classes, increases the probability of generalizing secret data. SVM offers the best classification performance i.e. the accuracy of the training set. It does not overflow the data.

SVM does not make strong assumptions about the data. Show more efficiency in the correct classification of future data. SVM is classified into two categories, i.e. Linear and non-linear. In a linear approach, training data is represented by a line, i.e. hyperplane, shown separately.

Consider the problem of separating the set of training vectors belonging to two distinct classes,

G = {(x_i; y_i); i = 1; 2;:::;N} with a hyperplane $w^{T*}(x)$ + b =0 (xi is the ith input vector, yi∈ {−1; 1} is known binary target), the original SVM classi1er satisfies the following conditions:

$$w^{T*}\phi(x_i) + b \geq 1 \text{ if } y_i = 1$$
$$w^{T*}\phi(x_i) + b \leq 1 \text{ if } y_i = -1 \quad \text{(iv)}$$

where ϕ: $R^n \rightarrow R^m$ is the feature map mapping the input space to a usually high dimensional feature space where the data points become linearly separable.

The distance of a point xi from the hyperplane is

$$d(x_i, w, b) = \frac{|w^T * \emptyset(x_i) + b|}{|w^2|} \quad \text{(v)}$$

The margin is 2/|w| according to its definition. Hence,we can find the hyperplane that optimally separates the data by solving the optimization problem:

$$min\emptyset(w) = \frac{1}{2}|w|^2 \quad \text{(vi)}$$

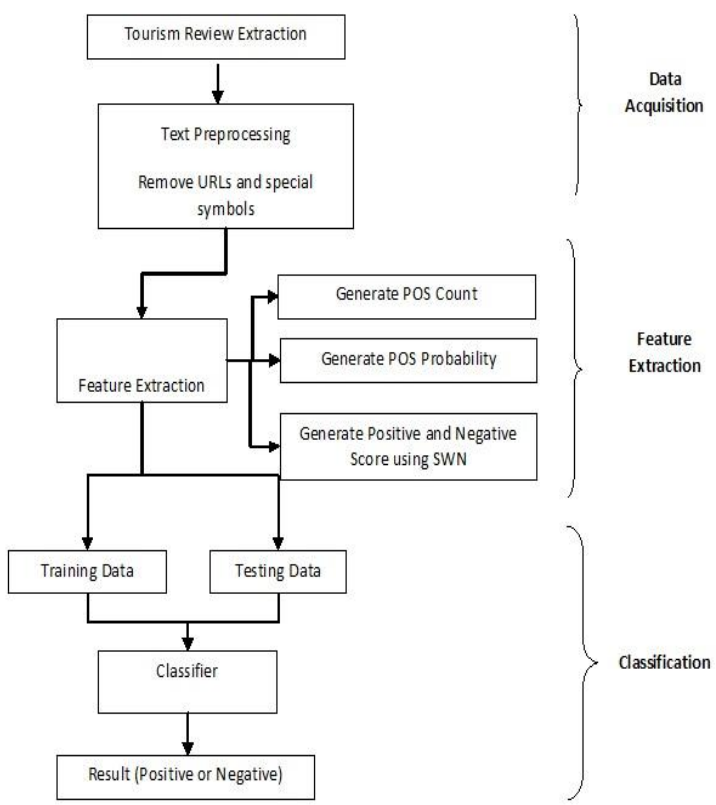

Figure 4: Proposed Flow Diagram of Tourism Review Sentiment Analysis

For the inseparable linear problem, we first assign the data to another large space H using a non-linear mapping, which we call Φ. So we use the linear model to achieve classification in new space H. Through defined "kernel function" k, is converted as follows:

$$max \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} a_i a_j y_i y_j k(\overrightarrow{x_i} * \overrightarrow{x_j}) \quad \text{(vii)}$$

$$s.t. \sum_{i=1}^{l} a_i y_i = 0 \qquad 0 \le a_i \le C, \quad i=1,2,\ldots\ldots l \quad \text{(viii)}$$

And corresponding classification decision function is converted as follows:

$$f(x) = sign\,[\sum_{i=1}^{l} a_i y_i k(\overrightarrow{x_i} * \vec{x}) + b] \quad \text{(ix)}$$

The selection of kernel function aims to take the place of inner product of basic function. The kernel function investigates the non-separable problems as follows:

$$k(x_i x_j) = \exp(-\gamma ||x_i - x_j||) \quad \text{(x)}$$

k-Nearest Neighbour (kNN)

kNN is used for both classification and regression problems. It is one of the simplest classification algorithms. Determine the parameter k which is number of nearest neighbours. When there is new data point to classify, then its k nearest neighbours is find out from the training data. The distance is calculated using one of the measure from Euclidean distance, Minkowski distance, Mahalanobis distance. The larger is k, the better is classification.

Random forest

Random forest is found as best model for prediction. It is learning method for classification, regression. Multiple decision trees are constructed at training time and outputting the classes or prediction. Random forest applies bootstrap aggregation technique which decorrelates the trees by showing them different training sets. For each tree, a subset of all the features can be used. As the number of decision tree increases, the variance of the model can be greatly lowered and Accuracy increases. In Random Forest, 2 main parameters are considered i.e. number of trees and number of features they select at each decision point. Accuracy of prediction increases as more number of trees making decisions. RF improves prediction accuracy as compared to single trees. RF handles larger numbers of predictors and it is faster to predict. RF found to overfit for some datasets with noisy classification tasks. Large number of trees may make the algorithm slow for real-time prediction [1,5].

The performance evaluation are performed using feature extraction technique using POS count and probability and SentiwordNet score. Table 3 shows the performance evaluation of different classification algorithm over datasets. From the result analysis it has been analyzed that accuracy of random forest classification achieved best result. Some of the performance parameters are discussed below:

Recognition Accuracy is represented as:

(TP+TN)/(TP+TN+FP+FN) (xi)

Precision is represented as:

(TP)/(TP+FP) (xii)

Recall is represented as:

(TP)/(TP+FN) (xiii)

F_measure is represented as:

(2*Recall*Precision)/(Recall + Precision) (xiv)

Where,

TP= True Positive, that means if one sample is anpositive and the predicted label also stands positive.

TN= True Negative, that means if one sample is negative and the predicted label stands negative.

FP = False Positive, that means if one sample is negative and the predicted label stands positive.

FN= False Negative, that means if one sample is positive and the predicted label stands negative.

TP stands the number of true positive samples, FN stands the number of false negative samples, FP stands the number of false positive samples, and TN stands the number of true negatives.

Table 3: Performance Evaluation of Proposed Algorithm

Algorithms	Accuracy	Precision	Recall	F_Measure
k-NN	80	76.7123	84.8485	80.5755
SVM	88.8889	86.3014	92.6471	89.3617
RF	93.3333	90.411	97.0588	93.617

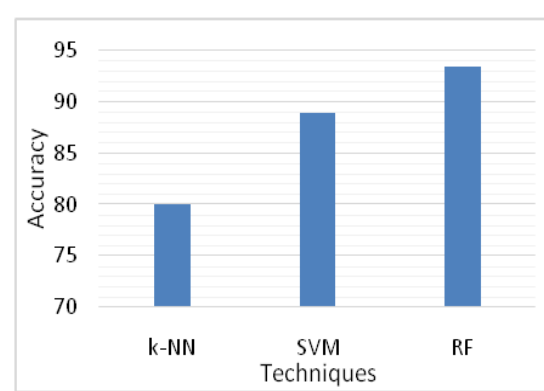

Figure 2: Performance Comparison of Accuracy

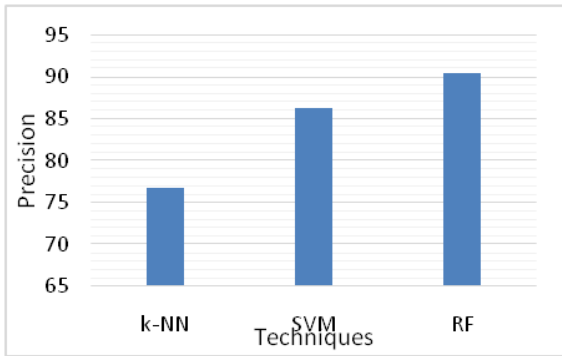

Figure 3: Performance Comparison of Precision

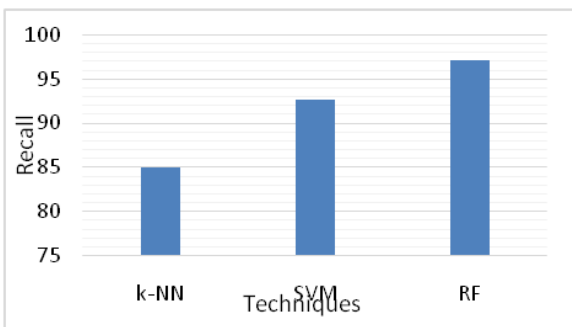

Figure 4: Performance Comparison of Recall

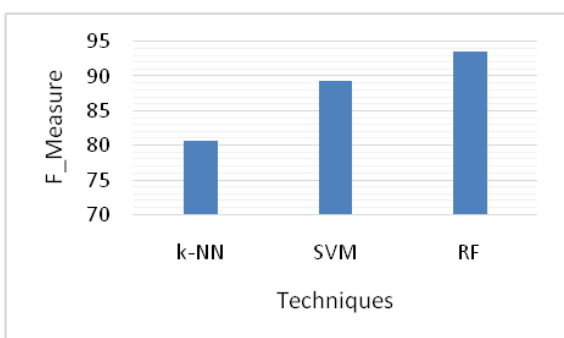

Figure 5: Performance Comparison of F_Measure

V. Conclusion

In this paper, lexicon-based sentiment analysis framework is proposed for tourist review with Part-Of-Speech (POS) tags in the lexicon. The training review sentiment scores are evaluated by the SentiWordNet (SWN). Then a lexicon-based sentiment analysis framework is proposed using machine learning approach for analysis of class of review either it is positive or negative. The proposed method improves the recognition performance of the review. The proposed algorithm is analyzed using three classifiers i.e. support vector machine (SVM), k nearest neighbour (KNN) and random forest (RF). After simulation results KNN gives about 80% accuracy, SVM gives about 89% accuracy and RF gives 93% accuracy.

References

1. Sruthi S, Reshma Sheik and Ansamma John, "Reduced Feature Based Sentiment Analysis on Movie Reviews Using Key Terms", IEEE, 2017.
2. Tirath Prasad Sahu and Sanjeev Ahuja, "Sentiment Analysis of Movie Reviews: A study on Feature Selection & Classification Algorithms", IEEE, 2016.
3. Md Shad Akhtar, Ayush Kumar, Asif Ekbal, Pushpak Bhattacharyya, "A Hybrid Deep Learning Architecture for Sentiment Analysis", International Conference on Computational Linguistics: Technical Papers, pp. 482–493, 2016.
4. K. M. Anil Kumar, N. Rajasimha, M Reddy, A. Rajanarayana, K. Nadgir, "Analysis of Users' Sentiments from Kannada Web Documents", International Conference on Communication Networks, vol. 54, pp. 247-256, 2015.
5. Namita Mittal, Basant Aggarwal, Garvit Chouhan, Nitin Bania, Prateek Pareek, "Sentiment Analysis of Hindi Review based on based on Negation and Discourse Relation", International Joint Conference on Natural Language Processing, pp 45-50, 2013.
6. Mohsen Farhadloo, Erik Rolland," Multi-Class Sentiment Analysis with Clustering and Score Representation", IEEE 13th International Conference on Data Mining Workshops, pp. 904-912, 2013.
7. Singh, V. K., et al. "Sentiment analysis of movie reviews: A new feature based heuristic for aspect-level sentiment classification." Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on. IEEE, 2013.
8. Amitava Das, Sivaji Bandopadaya, "SentiWordnet for Bangla", Knowledge Sharing Event -4: Task, Volume 2,2010.
9. Amitava Das, Sivaji Bandopadaya, "SentiWordnet for indian language", Workshop on Asian Language Resources, pp. 56-63, Beijing, China, 21-22 August 2010.
10. Aditya Joshi, Balamurali AR, Pushpak Bhattacharya, "A fall back strategy for sentiment analysis in hindi", International Conference on Natural Language Processing, 2010.
11. Annett, Michelle, and GrzegorzKondrak. "A comparison of sentiment analysis techniques: Polarizing movie blogs." Advances in artificial intelligence. Springer Berlin Heidelberg, 2008. 25-35.