" Tourists Behavior Prediction through Online Reviews by Analyzing their Sentiments using Machine Learning Approach"

Dr. Harsh Arora
Associate Professor
BCIIT,New Delhi
drharsh.cse@gmail.com

Abstract*— Sentiment Analysis is that part of research area where different people opinions or sentiments are extracted in form of textual data from various websites. In this paper sentiment analysis has been described along with machine learning techniques on tourist's reviews to see their behavior towards various tourist places, hotels and other services provided by tourism industry. Emotions in the form of tourist's reviews extracted are interpreted and classified by preprocessing of data and further feature extraction is done through machine learning highly efficient technique called deep learning. In this paper, the proposed idea has been given to use deep learning methods like CNN, RNN and LSTM rather than using machine learning classical algorithms like SVM, Naive Bayes, KNN, RF etc. Also, comparison of various machine learning and deep learning techniques working on tourist sentiments has been done here in this paper to show that deep learning techniques analyze and classify emotions and polarity with deep layers efficiently where on the other hand classical algorithms of machine learning give results not better than deep learning techniques. In this way sentiment analysis has been done and the proposed idea of this research paper is change in the machine learning techniques or methods from classical algorithms to neural network deep learning methods which in future definitely will give better results to analyze deeply the sentiments of tourists to find out the liking and disliking of various tourist places, hotels and related tourism services that will help tourism business industry to work on the gap in existing services provided by them and system can become more efficient in future. Such improved tourism system will give benefits to tourists or users in terms of better services and undoubtedly it will help tourism industry to enhance business in future.

Keywords—sentiment analysis, machine learning, deep learning, tourist reviews.

I. INTRODUCTION

Over the past few years, online tools have changed the travel and tourism industry. This is important to understand travel trends to provide easy and exciting travel experiences to the tourist in order to develop the new business model in the field of tourism industry. Data is being generated at very high rates and can be both structured and unstructured. The idea is to analyze tourist online reviews by sentiments classification various methods using machine learning. The key problem arises is that there is no known way of using data to generate values to know how to get benefit from these data. The four V's of data namely volume, variety, velocity and value irrelevant with the need for real time and customized information. The tourism industry is the industry where customer experience chill for the repetition and growth, has highly adapted to the evolving Technology. Age off automatic tools in social networks for tourism sector stated the importance of influencing the customer's involvement. It is affecting the way in which tourist perceive their experience and needs [1].Many tourist services are now available on the internet through online booking websites or portals. With addition travel is one of the dominating topics on social media. It is not surprising that tourism has been recognized as the number one sector in terms of online environment. The scores that are obtained by the predictive analytics from different model pictures that should be taken for the retainment of the client by offering the new services. Predicting behavior we need the reliable consistent and persistent information. It is timely to examine how tourism researchers are making use of these data new types of data form a part of a new research paradigm. In context of tourism a service based industry that relies on positive customer emotion and feedback of visitor satisfaction is of critical importance. Social interaction characteristic is the main focus of the digital conversation using new tools and methods. Electronic word of mouth is introduced feature which led to completely different consequences from traditional conversation. Using these, Google started creating their own content using web 2.0 technologies which focuses on the ease with which the message is easily distributed via social media. As a part of artificial intelligence, machine learning is a new Shine. By using machine learning algorithms on the big data we are enhancing the terms of tourism industry. Similarly, sentimental analysis technique has become a new verge.

II. METHODOLOGY

A. *Machine learning basic classification algorithms*

- Linear regression:

  Relationship between the input variable x and output variable y is expressed as an equation of the form y=a+ bx.

  Thus, the goal of linear regression is to find out the values of coefficient a and b. here a is the intercept and b is the slope of line.

- Classification and regression trees:

  They are one implementation of decision trees. The non-terminal nodes of c a r t are the root nodes and the internal nodes. Nodes are the leaf nodes. Non terminal node represent a single input variable x and a splitting point on that variable, the leaf nodes represent the output variable y. Is used as follows to make prediction the splits of the trees to arrive at a leaf node and output the value present at the leaf node.

- Naive Bayes:

  It is a technique based on an option of Independence among predictions using Bayes Theorem. Also known as generatic learning model. In simple words this classification technique assume that a particular feature in a class is not related any other feature present in the class. Even if all the features are dependent on each other, these properties independently contribute to the posterior probability. Name based classifier is very useful for enormous data sets and very easy to build. That is why it is used to outperform highly sophisticated classification method as well.To calculate the probability of hypothesis h being true, given our prior knowledge d, we use Bayes Theorem as follows:
  P(h/d)=(P(d/h).P(h)/P(d))where
  P(h/d)=posterior probability

  P(d/h)=likelihood the probability of data D given that the hypothesis is h was true.

  P(h)=class prior probability

  P(d)=predictor prior probability

- K-Nearest Neighbors:
  It is one of the simplest classification algorithm used for regression problem and classification problem. It basically takes a bunch of all available cases and then learns how to classify other new cases. This technique looks at all the available cases close to the new case by majority vote of it's K neighbors measured by distance function. This algorithm uses the entire data set as the training set rather than splitting the data set into a training set and test set. When an outcome is required for a new data instance, the KNN algorithm go through the entire data set to find k nearest neighbor tenses to the new instances or the k number of instances most similar to the new record and then output outcomes for regression problem for the Mode most frequent class for classification problem. The value of k is user-specified.

- Random Forest:

  Random forest is a method used for both classification as well as regression. This algorithm is operated by controlling decision trees samples and get the production from each of the trees and the best solution by means of majority voting. While constructing a tree at training time and outputting the class made up of predictions of the individual tree. It is better than single decision tree because it reduces the habit of over fitting to their training set.

- Support vector machine:

  It is a oversight machine learning algorithm used for classification and regression. It is majorly used in classification challenges. In this algorithm we plot each data item as a co-ordinate in n - dimensional space with the value of each attribute of particular point that is a co-ordinate. Support vector machine classifier executes the classification by finding the variance between two classes using hyper-plane or line.

- Neural network:

  Set of algorithms that is modeled loosely after human brain, design to recognize patterns. Basically, it is a process that makes the way a human brain operates according to the real world underlying relationships in a set of real world data. Refers to the system of neurons, original or artificial in nature. It interprets the sensory data through a machine like structure, clustering raw input or labeling that recognizes the numeric facts, all real world data be it text, sound, time series, images. Neural network help in classification and clustering which can adapt to changing inputs, then network generate the best result without altering the design of output criteria. It forms the unlabeled data according to the similarities among the input and then it classified the data to have a labelled data set to train on. Neural Network is a mathematical function that collects and then classify according to a designed architecture. So, you can think of neural network as a component of larger deep learning applications.

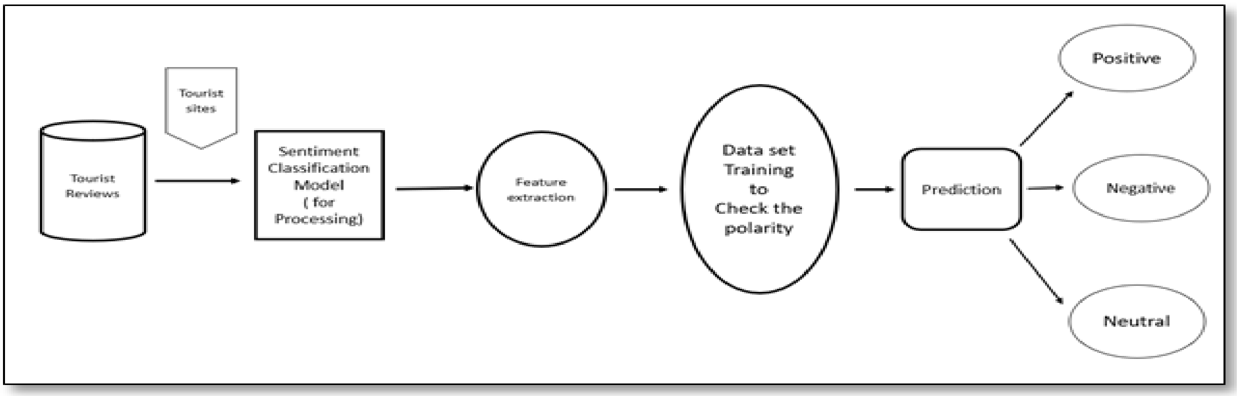
Tourist Reviews
Tourist sites
Sentiment Classification Model (for Processing)
Feature extraction
Data set Training to Check the polarity
Prediction
Positive
Negative
Neutral

Fig1: Tourist data set pre-processing and prediction

B. *Deep learning and its future*

Deep learning comes in the type category of machine learning and part of artificial intelligence. It works in the ways as like human brains and certain types of knowledge. It is an essential part of data science, which include statistics and predictive modelling. These days deep learning techniques work best on bulk of data taken through various websites in order to analyze the tourist reviews. In simple terms, we can think of deep learning as a way to automate predictive analytics. Algorithms of deep learning are stacked in a hierarchy of high to low complexity and abstraction. There is a non linear transformation of the input and uses what it learn to obtain a statistical model as output. As the process continues, number of layer from which the data passes through also increase which give the inspiration of its name deep learning.

Future of deep learning:

- Enhancement in set of standard tools: The end of this decade, the deep learning community will convert on a core set of de facto tooling frameworks.
- Deep learning will gain native support within spark:
  Spark community well Beef Up The platforms native deep learning capabilities.
- Deep learning will search a stable slot within the open analytics ecosystem:
  What's becoming Clear is that you can't adequately train, manage and deploy deep learning algorithm without the full suite of Big Data Analytics capabilities.
- Deep learning tool will incorporate simplified programming frameworks for fast coding: When moving forward, deep learning developers will endorse open, systematized, cloud based development environment.

Deep learning tool will get assist visual development of reusable components:
Deep learning tool kits will consolidate modular capabilities for easy visual design, configuration and training of new model from already existing building blocks[2].

C. *Comparision of machine learning and deep learning approach.*

Deep learning and machine learning are two widely known technologies in the world today. These two technologies are usually used interchangeably. More clearly, Deep Learning is the part of machine learning. Since, the deep learning is a subset of machine learning many of us get confused between both of them.

Data dependencies:

The main difference between both algorithms is a performance. So when the data is small, the deep learning algorithms don't perform well. That is why the deep learning algorithm needs a large amount of data.

Hardware dependencies:

Dependency of deep learning is on high end machines. While when we talk about traditional learning sweet depends on low end machine. The requirement of deep learning also includes GPUs.

Feature engineering:

In this process, the knowledge is creation of feature extractor. The complexities of the data are also reduced to some extent. Consumption of time is more.

Problem solving approach:

In general, we use traditional algorithm for solving the problem. What is needed, solar problem should be broken into different parts. Afterwards it will be solved individually.

Execution time:

Usually, there is more time consumption in deep learning as compared to the machine learning for the training. The reason for the long time consumption is that usage of many parameters in deep learning algorithm why machine learning takes much less time to train[3].

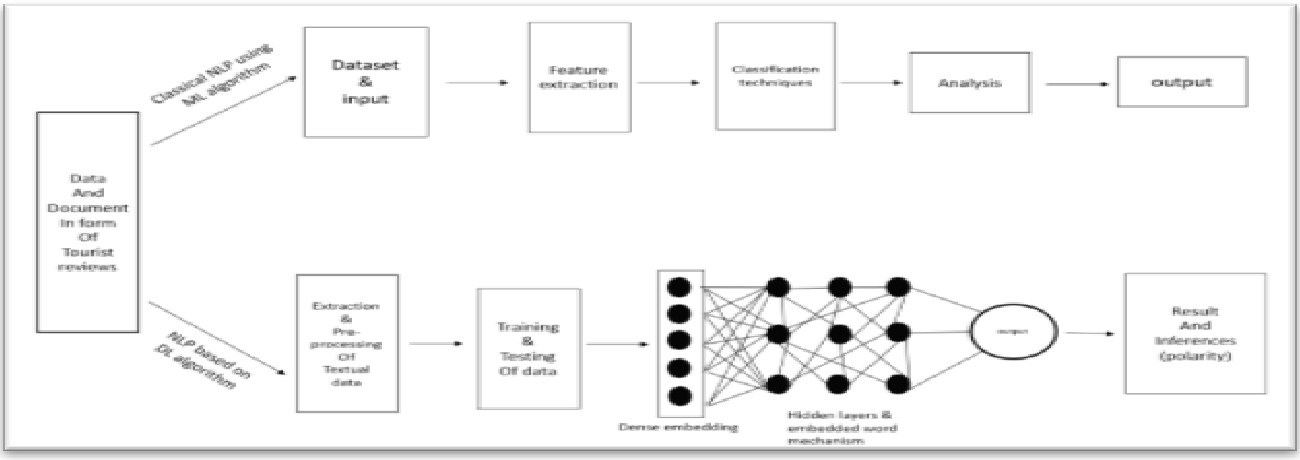

Fig2: Comparison of ML based NLP vs DL based NLP

III. ONLINE REVIEWS

Development of e-commerce has drastically changed the way that customer search for any kind of information or make their preferences regarding, or to purchase services are products. Customer taken by questions such as “find the best location?”. “where is the best hotel?”, “Is it a famous hotel ?”. “ Is it really worth to give money in this booking?”. The customer tries to search answers to these questions, consult online reviews posted by other customer who have appropriate and relevant experience. Online reviews are very crucial to both buyers and seller and also researchers focus on 3 point of online reviews, which are:

1. Negative
2. Positive
3. Neutral

And their effect on purchase intention. The valence (or negative) online review and rating has a outshine impact on potential customer and their decision of purchasing. The findings shows a positive correlation between the valence of reviews and the hotel purchasing intention. High rating can generate customer revisit intention. Previews and the proportion of good reviews, specially when the rating can have a great impact on restaurant price hotel rates as well.

IV. WHY DEEP LEARNING

The base of the tourism industry is a services that include travel, transportation, accommodation and similar services. Everyday billions of dollars changes hands in the complex ecosystem. It is that kind of industry where the customers wishes meet more than their needs. Machine learning and deep learning can improve the strategies of competitors to offer a better service.

- Prediction of seasonal demands for services:

  Tourism is a business which is marked by the seasonality of demand. At the peak season, tourism Product Suppliers have the opportunity to earn more. It is easy and accurate to find the correlation between the using deep learning algorithm that cause this seasonal demand by analyzing the raw data and protect the Trends for future. This is called predictive analytics on time series.

- Pricing strategies:

  One of the major strategies of the tourism Product Suppliers is to attract the customer by using competitive prices. Companies adjust the price without compromising their profits. Here, deep learning prove to be useful by analysing the data like history of Hotel, local events, competitions or promotions. These data can be analyzed through predictive model which provides best possible prices.

- Personalized recommendations:

  For a long time, many known travel sites have used recommendation engines. Travel sites offers their user the holiday packages which best fit their customers profile. These engines collect the info like preferences, budget related data, and customers detail to give the customer a personalized travel recommendation. Information acquired is used to find best possible alternative by comparing the option.

- Customer experience:

  In customers profile, there is a large variation as their demands and expectations are different. Every customer want that they should be treated according to their preferences. So, area of the sector apply market segmentation. The entire chain of customer is divided into segments and then subdivided on the factor of similar characteristic, demand and expectation. By this process, customer can be offered a much more personalized and specialized service[4].

V. SENTIMENT ANALYSIS OF TOURIST REVIEWS USING DEEP LEARNING TECHNIQUES.

Sentiment analysis:

It is also known as opinion mining which is the analysis of the feeling (attitudes emotions and opinions) words using Natural Language Processing tools. It's looking beyond the number of likes shares are comments you get on an ad campaign, product release, blog post and video to understand how people are responding to it. What's the review positive? Negative? Sarcastic? ideologically biased? Opinion target, opinion holder and opinion are the definition used to extracting opinion from different online sources. An opinion can be expressed in two types and these are Direct opinion and Comparative opinion. All the opinion are stored in a document. Following are the steps to extracting the opinions:

1.Identify the object
2.Feature extraction and synonym grouping opinion orientation determination
3Integration

Sentiment analysis comprises a multi-step process:
1) Data Retrieval
2) Data Extraction
3) Data Preprocessing
4) Feature Extraction
5) Topic Detection
6) Data Mining Process

Data retrieval requires the identification and definition of the data source. To collect the review data from these sources a specific web crawling mechanism is necessary to fetch the data and then save them in a database. After collecting data in a database view data needs to be extracted then set of heterogeneous data fields. The review text needs to be extricating using appropriate expressions. Is extricate review contains one or several sentences impact the reviews opinion. Part of speech tagging is an important preprocessing task .It is a part of sentiment analysis by appointing each word a particular label. Feature extraction is known as the process of pulling out a set of discriminative informative and non-similar values to numerically represent a review or text[5].

5.1. *Convolutional Neural Networks.*

Another type of neural network that can be used to predict time series is a convolutional neural network (CNN). These are biologically inspired variants of feed-forward neural networks used primarily in computer vision problems although their ability to exploit spatially local correlation in images can also be used in time-series problems, like sentiment analysis. In these models, the output of each neuron is generated from the output of a subset of spatially adjacent neurons. Every neuron in the same layer shares the same weight and bias, meaning the layer can be expressed in terms of a filter that is convoluted with the output of the previous layer. In order to apply a CNN to a time series, we arranged the encoded words in the same order as they appear in the comment, such that adjacent words in the comment are spatially adjacent at the input to the neural network. Moreover, each word embedding dimension is a different input channel to the network. In this way, the convolutional layers can exploit the local correlation between words in each comment. After each convolutional layer with a ReLU, activation function is a max-pooling layer, which partitions the input into a set of non overlapping ranges and, for each range, outputs the maximum value. Following the convolutional and max-pooling layers is a feed forward layer to yield the output of the entire network.

5.2. Recurrent Neural Networks.

To predict the sentiment of the comments, we use models based on neural networks. Each comment is a sequence of encoded words that can be processed as a time series. However, the most common neural networks (e.g., feed-forward neural networks) lack the memory to store information over time. Recurrent neural networks (RNN) solve this problem by making the network output *yj* at step *j* depend on previous computations through a hidden state *sj* that acts as a memory for the network.

Figure 3 shows the RNN we used, unfolded into a full network. By unfolded, we simply mean that we write out the network for a complete sequence of *N* steps, where *xj* is the *j-th* encoded word in the comment, which we used as the input to the network in the j^{th} step.

In a RNN, the relationship between output *yj*, input *xj*, and state *sj* in step *j* is determined by the type of RNN cell.

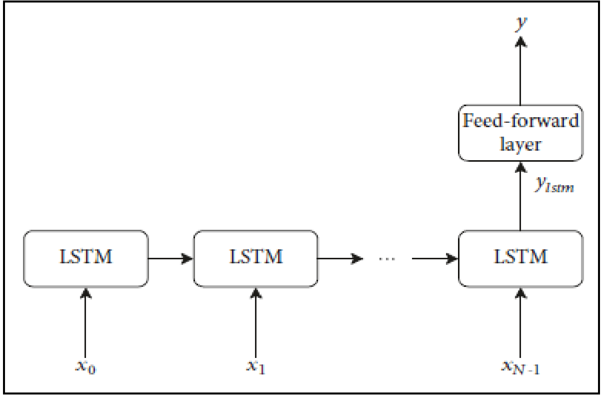

Fig3: LSTM unfolded into a full network.

As Figure 3 shows, we used a kind of cell called long short-term memory (LSTM). LSTM recurrent neural networks are capable of learning and remembering over long input sequences and tend to work very well for labelling sequences of word problems. As (1) shows, output y_j depends on the state s_j of the LSTM cell through the activation function $\sigma_y(x)$ (which is Generally, tanh (x)). The state s_j depends on the state of the previous step S_{j-1} and on the candidate for the new value ofthe state s_j. The output gate o_j controls the extent to which the state s_j is used to compute the output y_j by means of the Hadamard product. The input gate i_j controls the extent to which s_j flows into the memory, and the forget gate f_j controls the extent to which s_{j-1} remains in memory. The o_j, f_j, and i_j gates and the candidate for the new value of the state of the cell s_j can be interpreted as the outputs of conventional artificial neurons whose inputs are the input to cell x_j at step j and the output of cell y_{j-1} at j − 1. The activation function for the gates σ_g is usually the sigmoid function, while for σ_s it is usually tanh (x).

$$\begin{aligned}
f_j &= \sigma_g\left(W_f x_j + U_f y_{j-1} + b_f\right),\\
i_j &= \sigma_g\left(W_i x_j + U_i y_{j-1} + b_i\right),\\
o_j &= \sigma_g\left(W_o x_j + U_o y_{j-1} + b_o\right), \qquad (1)\\
\tilde{s}_j &= \sigma_s\left(W_s x_j + U_s y_{j-1} + b_s\right),\\
s_j &= f_j \circ s_{j-1} + i_j \circ \tilde{s}_j,\\
y_j &= o_j \circ \sigma_y\left(s_j\right).
\end{aligned}$$

As Figure 3 shows, each input x_j in (1) is the coded value of the successively coded sequence of words $(x_j)^{N-1}$ at j=0 in the comment. The RNN cell provides an output at each step j, but for the prediction, only output y_{N-1} of the network is considered when the last word x_{N-1} is input into the network.

This output is what we refer to as y_{lstm} in Figure 3.

Output y_{lstm} is used as an input to a one-neuron feed forward layer with a sigmoid activation function, the output of which, between 0 and 1, is the network's prediction for whether the sentiment is positive or negative. The output y of that single neuron can be expressed as indicated in

$$y = \sigma_{\text{sigmoid}}\left(\sum_k w_k y_{\text{lstm}} + b\right)\sigma_{\text{sigmoid}}(x) = \frac{e^x}{e^x + 1}, \qquad (2)$$

where w_k is used to denote the weight of the *k-th* input, *b* is the bias, and $\sigma_{sigmoid}$ (x) is the sigmoid activation function of the output neuron in the layer[6].

VI. RESULT AND DISCUSSION

It has been observed by analysing results though various studies and research done earlier in the field of tourism to check the emotions of tourists using classical algorithms of machine learning that accuracy measured by the same and by deep learning methods differs remarkably. Results of deep learning techniques are much better than classical techniques of machine learning. Deep learning techniques like CNN, RNN and LSTM will be applied in future as they all work on hidden layers and accordingly data set will trained and tested. The research will be focussed on feature extraction for sentiment analysis from tourist reviews. In order to find the simulation results, execution will be done on the data set taken from tourist sites to get to know the tourist various opinions. After that data will be fed to CNN, RNN or LSTM technique in order to work on the deep and hidden layers of data to find out the best results. Accuracy will be analysed and will be compared with accuracy of classical techniques of machine learning. In other words, performance comparison of accuracy of different machine learning algorithms will be done to compare it with one of the deep learning techniques to find the better efficiency.

VI I. CONCLUSION

In this paper, deep learning techniques CNN and RNN have been proposed to work on tourist review data. The proposed framework of sentiment classification for tourist reviews is to find out positive, negative and neutral reviews and the corresponding methodology has been mentioned in this paper. Further, comparison has been done of various classical and neural network techniques to find out which one is the better technique for future. More clarity of tourist sentiments, better will be the technique. Proposed techniques will give better results rather than existing work based on classical algorithms like SVM, KNN etc. Future work can be perused in several directions. This paper worked on the limitations of existing classical machine learning methods and proposed better techniques of deep learning to give far better accuracy of performance.

REFERENCES

[1] Aurchana.P,PIyyappan.R ,PPeriyasamy.P, "Sentiment Analysis in Tourism Dept. of M.CA, Sri Manakula Vinayagar Engineering College, Puducherry, November 2014 .*(references)*

[2] InfoWorld, "Extreme analytics,"[Online]. Available: https://www.infoworld.com/article/3172554/6-predictions-for-the-future-of-deep-learning.html

[3] +DataFlair,"deep learning vs machine learning."[Online] Available: https://data-flair.training/blogs/deep-learning-vs-machine-learning/

[4] OnlineTravelTechnology,"deep learning,"[Online] Available: https://onlinetraveltechnology.com/en/what-is-deep-learning-and-how-does-it-apply-to-tourism-technology/

[5] Alireza Alaei1, 2, Susanne Becken2, Bela Stantic1, Sentiment analysis in tourism: Capitalising on Big Data, 2Griffith University, 4222, Australia.*(references)*

[6] C. A. Martín , J. M. Torres , R. M. Aguilar , and S. Diaz "Using Deep Learning to Predict Sentiments: Case Study in Tourism" Available: