JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

TEXT MINING AND SENTIMENT ANALYSIS OF TRIP ADVISOR REVIEWS

Chandrashekar D.K, Aditi Bansal, Akanksha, Ruchi Kumari, Udaya Shankar K R

Associate Professor Student Student Student Student

Computer Science and Engineering,
Cambridge Institute of Technology, Bengaluru, India

Abstract: In the world of growing technology, each travel website produces a voluminous amount of data. These data are large in volume, velocity, variety, and veracity. It is humanly impossible to read and analyze each review manually. Therefore, the need for semantic analysis of trip advisor reviews. Trip Advisor is a web 2.0 application that allows the customer to share their experience with hotels, restaurants, and other tourist spots. Trip advisor is among the top traveling websites. Every day, millions of users visit the website while planning their trip. It is based on these reviews and ratings; trip advisor suggests a place for new users, but the number of reviews is in the thousands, and it requires sentiment analysis of all the existing reviews to move toward a conclusion. The opinion and view expressed by a reviewer in their review intuitively represent their evaluation of a destination. Their words represent a strong emotion which is important for both, future tourists and service providers. The classic sentiment analysis, which is predominantly centered on statistical data, has certain drawbacks since it is comparatively difficult to evaluate the overall feelings of a text, and is less effective in tasks involving Natural Language Processing (NLP). As a result, we employ aspect-based sentiment analysis (SA) based on the Transformers' Bidirectional Encoder Representation (BERT) model.

IndexTerms **- Artificial Intelligence, Big Data, BERT, Cloud, NLP, Text Mining, Sentiment Analysis**

I. INTRODUCTION

Word Of Mouth (WOM) has a great impact on people's behavior toward a thing. It greatly influences how people judge and react toward a particular product or service. People are more inclined towards buying already experimented products. Reviews and ratings have a huge impact on deliverable sales. Therefore, even in the case of tourism, travelers seek the help of online reviews before making decision about a hotel or tourist destination.

To evaluate the conclusion of thousands of reviews, we need to do a sentiment analysis of these reviews using Natural Language Processing (NLP). NLP is a sub-section of Artificial Intelligence that is concerned with human languages. All the web-based task or voice assistants such as Siri or Alexa or any other chat bots, all of these services use NLP to understand what human says and works accordingly.

As of now, sentiment analysis can be completed in two ways which is statistical-based and deep learning-based sentiment analysis approach. Statistical based sentiment analysis method is one where certain words or phrases determine the sentiment of the entire document. But in this paper, we will use a deep learning-based sentiment analysis model which is constructed on the ground of neural network. This model allows the network to predict the next word in the series based on the context of previous words. A few of the available deep learning models are Convolutional Neural Networks (CNN), recurrent neural networks (RNN), long-short-term memory networks (LSTM), and Transformers, etc. In this paper, we will use the Bidirectional Encoder Representation from the Transformers (BERT) model which is different from all the previously existing models and allows the transformer to read in both directions.

BERT model has an encoder stack of transformer architecture which is different from traditional encoder-decoder architecture. The BERT model is trained using a large dataset which facilitates easy prediction of the next word or phase in the row. The BERT model helps to understand public opinion about a particular product or service whether it is positive, negative, or neutral. It helps the company or the service provider to gain +++public opinion on their product and further put efforts to improvise it. Using the BERT model for sentiment analysis of reviews posted by the public is an important innovation of this paper. Another important innovation is the concept of cloud computing which allows parallel computing of tasks thus reducing the total time and cost invested in carrying out the sentiment analysis of the tourism reviews. This paper aims to increase code optimization, reduce memory spacing, and use cloud services for the best output. This paper aims to bridge the gap between previous research papers and current technology by using the BERT model.

II. RELATED WORKS

A large number of international publications is a valuable asset for tasks such as research and development. A literature review provides an overview of any technical or non-technical project and valuable information. The research in this domain is focused on extracting big data information from the digital world using various technologies combined with cloud computing and training it to obtain a sentiment analysis of TripAdvisor reviews.

Hossein Hassani *et.al.,* [1] implemented text mining, which is a method used by researchers, scientists, and interested candidates to maximize the value of the vast amount of text available digitally. It necessitates the application of text mining technologies, which have grown in popularity in recent years. It has applications such as the categorization of sentiments and opinion mining, which forecasts someone's opinion based on the media messages that were exchanged. It also enters the field of argument extraction, where arguments and facts are taken from political speeches and publications using text mining. Blog mining is a technical subset of text mining that enables authors to effectively interact with their readership. There are many different types of blogs, as well as both personal and business blogs. It resembles email mining and social media mining. Email data is noisy compared to text data; hence email mining has some aspects that are different from text mining. Also, there are differences in the format.

Ravi Vatrapu *et.al.,* [2] work specifies that text analysis, social network analysis, social complexity analysis, and social simulations are some of the major paradigms in the field of computational social science, which is where social media mining is carried out. However, each of these has some drawbacks that are overcome by social set analysis, a type of big data analytics. Empirical research on large-scale social data, which is used in cloud computing, serves as a representation of it.

Muhammad Inaam Ul Haq *et.al.,* [3] demonstrates that in the realm of cloud computing, information is extracted using techniques including term frequency analysis, similarity analysis, topic modelling (LDA), and cluster analysis, coupled with the application of text mining to big data processing. All of these methods identify the connections among the words in a corpus.

Alexandra L'Heureux *et.al.,* [4] research work shows that its common knowledge that machine learning assists with big data for cloud computing solutions to address issues relating to the volume, velocity, variety, and veracity of big data. The survey report by Alexandra L'Heureux et al. underlines the different problems of machine learning and links them to the aforementioned traits. It also identifies ways used to address the challenges that are likely to arise. The numerous tactics included ML criterion adaption, big data algorithm manipulation, and modifying existing models by the required algorithm.

Jingyi Zhang *et.al.,* [5] designed a study that shows how big data processing, when compared to conventional processing methods, carries out an assessment with higher accuracy and can be applied to evaluate tasks which have a high potential for risk challenges. This is due to the growing popularity and development of computing technologies like that of the Internet along with cloud computing and IoT. It can be used to improve the development of the healthcare sector as a whole and sports health management. Several wearable mobile medical devices and the utilization of genomic technologies make use of the Motion Risk Assessment method found in Big Data Analysis. In the survey paper, neural networks are used to analyze risk assessments related to major sports.

Mingchen Feng *et.al.,* [6] implemented cutting-edge data mining and deep learning techniques, big data analytics and mining that were applied to criminal data. In comparison to neural network models, the Prophet Model and LSTM outperform them, according to the data. To produce comparably superior trend prediction in terms of RMSE and spearman correlation, the study was used to examine criminal activities from three US cities for the ideal time of 3 years for the training sample.

Xianghua Fu *et.al.,* [7] propose sentiment analysis that uses LSTM to extract precise information from unstructured data and categorize it into positive or negative sentiment classes. The accepted procedure is to teach word embeddings and text representation to LSTMs. Word embeddings should not be used for sentiment analysis activities. To acquire the sentiment embedding of every word in the text, the survey paper has developed a lexicon-enhanced LSTM model, which offers findings that are comparably accurate.

Zhi Li *et.al.,* [8] study shows how to analyze danmaku reviews for sentiment using techniques like sentiment dictionaries and Naive Bayes (SD-NB), which helps monitor a danmaku video's sentiment factor and forecasts its rising popularity.

Yassin S. Mehanna et.al., [9] suggested a semantic conceptualization technique that takes into account all information embedded in the context to shorten it, to determine the proper sentiment towards the real target thing. This method uses tagged bags of concepts for SA. Tagged bag-of-concepts (TBoC) is a unique method for dissecting text to analyze the sentiments that are invisible meanwhile keeping all relationships intact and thus increasing the sentiment analysis's precision.

Wenjuan Luo *et.al.,* [10] formulated a study in contrast to the bag-of-words formulation of research articles which deploys the method by creating a conceptual framework on Latent Dirichlet Allocation with oblique monitoring that employs the 4×4 head, modifier, rating, and entity approach to establish the relationships connecting modifiers and ratings.

Staphord Bengesi *et.al.,* [11] conducted two stages of a study on a monkeypox outbreak employing SA. Using VADER and TextBlob, sentiment analysis was performed on over 500,000 multilingual tweets relating to the monkeypox post on Twitter in the first stage to categorize tweets by annotating them into positive, negative, and neutral feelings. 56 categorization models were designed, developed, and evaluated as part of the second stage of the project. The standardization of the vocabulary involved the use of stemming and lemmatization techniques. Vectorization was accomplished using learning techniques such CountVectorizer and TF-IDF methods, K-Nearest Neighbor (KNN), Logistic Regression, Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest, XGBoost and Naive Bayes. Recall, Precision, F1 Score and Accuracy were used to determine the functioning. The trial findings revealed that the depiction with the highest accuracy of around 0.935 was created utilizing SVM, Lemmatization, TextBlob annotation and CountVectorizer.

Rachmawan Adi Laksono *et.al.,* [12] analyzed reviews in internet media like travel advisor that do well based on consumer behavior. This study uses the Naive Bayes method to categorize restaurant patron satisfaction in Surabaya. Using WebHarvy Tools, data sampling is being done here through crawling. The results demonstrate that the Naive Bayes method, which has an accuracy of 2.9%, performs better than TextBlob sentiment analysis in obtaining the consumer reaction accurately.

Ren Cai *et.al.*, [13] explained BERT-BiLSTM which stands for Bidirectional Encoder Representations from Transformer Bidirectional Long Short-Term Memory is the amalgamation way of predicting Bidirectional Encoder Representations from Transformer (BERT) with Bidirectional Long Short-Term Memory, was used by Ren Cai et al. to classify the sentiment orientation (BiLSTM). It contrasts with BERT and BiLSTM models and predicts the sentiment orientation of consumer opinions. BERT-BiLSTM model's accuracy and recall are 0.8620 and 0.7078, respectively, exceeding those of the BERT model's 0.8559 and 0.5576 and the LSTM model's 0.7775 and 0.0747.

Yiren chen *et.al.*, [14] specified in their study that by adjusting the small datasets, self-supervised attention (SSA) keeps BERT from overfitting them. A hybrid strategy was suggested to incorporate the advantages of two alternative approaches to SSA integration into BERT. On all datasets, the hybrid model SSA-Hybrid navigates more effectively. It raises the base BERT's average score from 78.4 to 79.3.

Yan Cheng *et.al.*, [15] published paper that introduces the Multi-channel Convolution and Bidirectional GRU Multi-Head Attention Capsule that employs capsules to reflect the expressions contained in the text instead of scalar neurons and then replaces vector neurons with those. Multihead attention picks up term interactions and sentiment utterances that have been woven into the message. In order to retrieve both regional and global semantic elements of text, gated recurrent unit networks (Bi-GRU) with convolution neural networks (CNN) are used.

III. OBJECTIVES

The objectives of this project are to help the users in making quick decisive decisions on the trips and places they want to go on or visit. To identify the frequency of usage of each word through which key featured words can be extracted. To understand user's emotions based on their reviews using sentiment analysis. To curate the outcome to help the tourist industry by increasing their efficiency of marketing and their SEO scores. To cater user specific tourist attractions suggestions based on their preferences derived from their reviews.

IV. METHODOLOGY

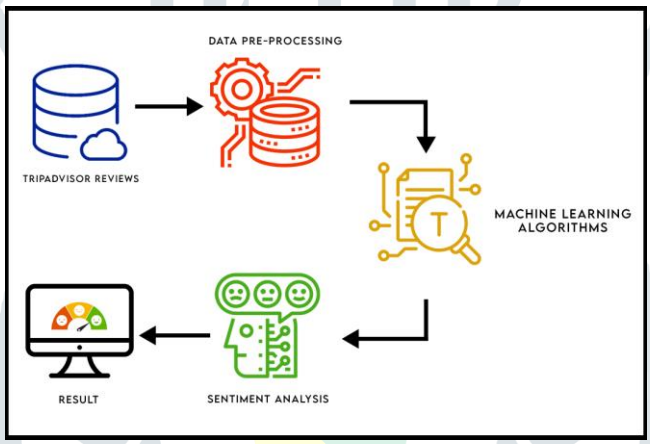

Figure 4.1: Text Mining and Sentiment Analysis of Trip Advisor Reviews

Trip Advisor Reviews

The process starts with accumulating dataset for the analysis. Data aggregation is the method of accumulating and analyzing data on favored variables in a standardized manner to address research aim and objectives and issues, test hypotheses, and assess conclusions. Figure 4.1 shows the different processes that goes in the analysis of Trip Advisor Reviews. All academic disciplines, including the humanities, social sciences, business, and basic and practical sciences, follow the same framework for gathering data. Although the approaches differ depending on the job, precise and honest collecting is still necessary. The consistency of research depends on the accumulation of reliable data, irrespective of the topic being examined or the method chosen for defining data (qualitative, quantitative). Errors are much less inclined to occur when the appropriate data gathering tools are used. We will acquire a shortlist 1,000 attractions from online, and we will then utilize the names on this list to search for 2,000 reviews (https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews) among these attractions on tripadvisor.com. The small amounts of information in this dataset will make the conclusions drawn from it less reliable because training neural networks requires a big quantity of data. An iterative translation-based text enhancement technique will be used to enrich the initial dataset and raise the dataset's 2,000 System Design evaluations to 4,000 reviews to improvise the experiment's generalizability and authenticity. As a technique of enhancing the data in this case, sentence generators (SG) and sentence discriminators (SD) based on GAN will be used. The two are taught independently, with SD being more analogous to a text classifier and SG seeking to deliver comprehensive augmented expression utilizing serial and parallel rotations of previous translators. In order to select out mediocre sentences using SD, SG must learn to provide high-quality changed utterances when it can. In order to produce high-quality datasets, the two work together.

Table 4.1: Trip Advisor Dataset

TEXT	RATING	TAG
Impressive ambience, services were quick and up to the mark. Hotel is located at the center of the city. Also, some of the greatest tourist spots are just at the walking distance make it an obvious choice. Sizes of the room were bit small but manageable. Hotel provides dining, transportation, spa and many other amenities.	4	Positive
Hotel wasn't very great but not that bad. Good location, clean rooms, size was small, overpriced food and drinks, noisy, staffs and services were good. Toiletries and other products provided were of high quality and smelled amazing.	3	Neutral
One of the worst stay I have ever had. The hotel was highly overpriced, and services provided were so disappointing in such high price. We kept asking housekeeping to clean the room buy they took more than 2 hours to attend us. Also, food tasted bland, and drinks were watery	1	Negative
Absolutely in love with this property. I had an amazing and relaxing stay. All the services were just perfect. The hotel staffs were friendly and available all the time for any assistance. Also, the services were very quick. They have a lot of options, and the rooms were luxurious. Also, the ambience of the hotel and designing was very royal and beautiful. It's worth every penny spent for the stay.	5	Positive
Had a bad and disappointing experience. I had high hopes with the hotel, but it wasn't up to the mark. Services were slow. Overhyped food. Spa services very expensive. Also, our air conditioner stopped working suddenly and had to shift to another room. Parking area is small and mostly full. Valets are friendly though.	2	Negative

The Table 4.1 shown above contains Trip Advisor dataset. The first column comprises of reviews posted by users. The second column consists of ratings. Here rating specifies-
1 – worst, 2- bad, 3- average ,4- good, 5- excellent.
The third column consists of tag based on reviews and ratings.
Positive – highly recommend
Neutral- can be considered as an option.
Negative – not recommended.

Data Pre-processing

Data processing is required to refine the data that has been acquired. Data preparation is a crucial step in the data mining process. It describes the steps involved in cleaning, transforming, and combining data in order to prepare it for analysis. Data preprocessing is done to better the quality and applicability of the data for the data-mining function. As shown in Figure 4.1, it takes place after acquiring the dataset.
Some common steps in data preprocessing include:
Data cleaning: This procedure finds and eliminates any missing, inconsistent, or irrelevant data. This might encompass dealing with outliers, removing duplicated data, and adding numbers when they are unavailable.
Data integration: the operation of integrating data from multiple sources, such as databases, spreadsheets, and text files. The goal of integration is to create a single, uniform visualization of the information.
Data transformation: is the procedure for altering the data's format so that it is more compatible with data mining. This might mean creating dummy variables, standardizing numerical data, and converting category data to numeric values.
Data reduction: the method used to identify a subset of data that is relevant to the data mining activity. This may be done through feature selection (selecting a subgroup of such variables) or feature extraction.
Data discretization: the transformation of consistent numerical data into categorical data, which may ultimately be used in decision trees as well as other categorical data mining frameworks.
These procedures increase the effectiveness of data mining and improve the precision of the findings.

Machine Learning Algorithms

KNN (K-Nearest Neighbour)

A machine learning method called K-Nearest Neighbor is based on the approach of supervised learning. By implying similarity between the data instances and existing cases, the KNN model allocates a new case to the classification that most closely matches the other comprehensive program. A new instance is classified using the K-NN method depending on how similar prior instances are to the current one. This implies that if a new data instance arises, it may be quickly classified using the K-NN algorithm into an ideal grouping. The K-NN algorithm may be used to solve classification and regression issues, but it is most frequently employed to solve classification problems. K-NN is a non-parametric method since it does not make any assumptions about the underlying data. Because it saves the training set's information during classification rather than learning from it instantly, the method often known as lazy learner. As depicted in Figure 4.2, in the training stage, the KNN model only saves the dataset as and when its new data input is generated, it categorizes the stored data into a cluster that almost resembles the new data. We must choose the Kth Neighbour in order to build a KNN model.

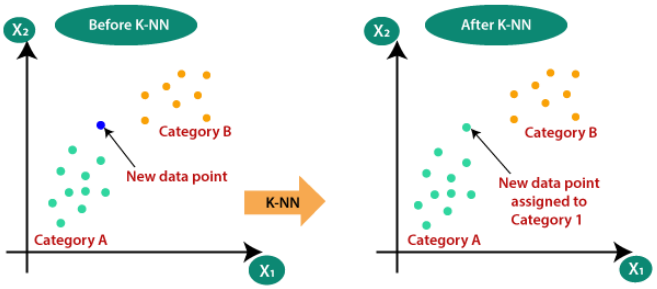

Figure 4.2: Dataset class before and after K-NN

Finally, we place the new data points in the category where the neighbor count is at its highest. In the case of sentiment analysis, we may also employ a similar strategy. A text's emotion can be used to categorize it, providing a rough categorization. There are still some drawbacks to this strategy. The calculation of the proper value of K is computationally challenging. As K plays a crucial role in categorization, determining the right value for K can be challenging. Moreover, KNN struggles to build classes from huge datasets due to the increased noise. And since we are dealing with Text classification, the datasets are generally large, hence creating this issue. The datasets are typically enormous because we are dealing with text categorization, which contributes to this problem. Additionally, the dataset for text classification, particularly for sentiment classification, contains a large number of noisy words because people frequently include words in reviews that have nothing to do with their sentiment, such as a nearby tourist attraction, which makes it challenging for our model to make accurate predictions.

SVM (Support Vector Machine)

One well-liked approach for supervising learning is the support vector machine, which is useful for both classification and regression issues. In machine learning, SVM is used to solve classification problems. Building a decision boundary that can categorize an n-dimensional space is the primary objective of the SVM method .

The hyperplane decision boundary carries this name. SVM chooses the corner points that aid in the hyperplane's construction. As these corner point examples are referred to as support vectors, this approach is also known as a support vector machine, or SVM. Two distinct categories that are categorized using a hyperplane are shown in the Figure 4.3. SVM is used to split the data into linear SVM and nonlinear SVM categories.

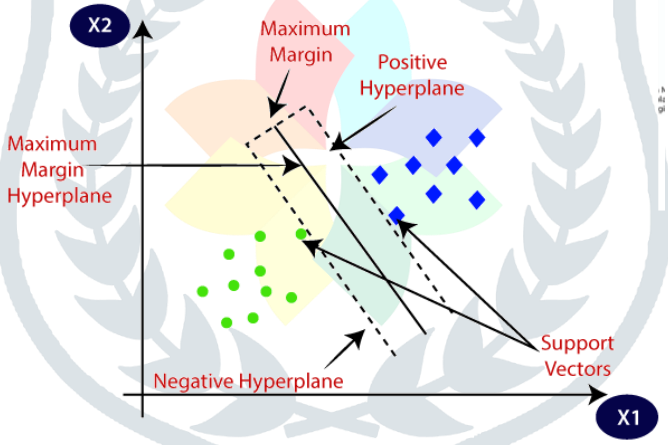

Figure 4.3: Hyperplane decision boundary of SVM

If we encounter a strange dog with characteristics of a cat and want to determine if it is a dog or a cat, we may use the SVM algorithm to accomplish so. We start by gathering several pictures of cats and dogs and studying their attributes. Then, we compare these features to our model (that weird dog). When using Corner Point as the support vector, a border between the dog and cat data is created. The support vector determined its classification as a dog.

Naive Bayes Classifier

It is one of the majorly often used classification methods in the early twenty-first century is the naive Bayes classifier. It is mostly used to determine the hypothesis' probability value. Which probability will hold depends on which has the highest likelihood.

Using the other hypothesis' likelihood of being true in the database, nave Bayes is one of the extremely reliable methods for identifying the most likely hypothesis. It uses the concept of Bayes theorem which says –

$$\Pi(A \square B) = (\Pi(B \square A).\ \Pi(A))/(\Pi(B)) \quad (1)$$

According to this formula, the likelihood of action X happening given that action Y has already happened is the exact same to the likelihood of action Y Happening given that action X has already happened multiplied by the likelihood of event X and divided by the likelihood of event Y. We employ a similar idea in the case of a Naive Bayes Classifier, where the B might be any or all of the hypotheses in the dataset. And as a result, this equation was later changed.

$$P(h \mid (x1, x2, x3 \ldots xn)) = (P(x1, x2, x3 \ldots xn \mid h).\ P(h))/(P(x1, x2, x3 \ldots xn)) \quad (2)$$

According to the aforementioned equation, we may estimate the cumulative probability that each of the n cases in the dataset, where x1 represents the first instance, x2 the second, and so on, is true. Also, we are aware of the total likelihood that these occurrences will

occur if 'h' already exists. Hence, assuming the other events are already true, we can determine the probability that it will occur using the Bayes theorem. Hence, using this theorem, we can determine the likelihood that a consumer would enjoy a certain hotel, given that we already know the likelihood that other past customers have either loved or hated the product. This process could be easily used for the task of sentiment analysis hence Naïve Bayes theorem becomes a key source in Sentiment Analysis.

BERT (Bidirectional Encoder Representations from transformers)

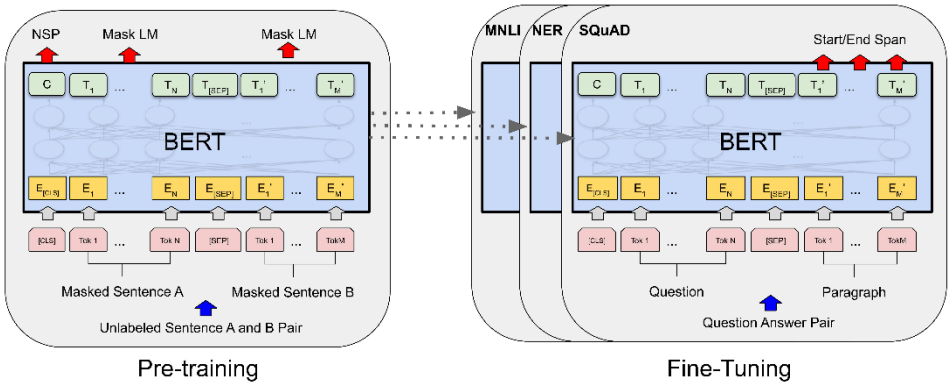

Figure 4.4: Pre-trained language model BERT with Fine-Tuning

The layers of a transformer encoder that makes up the BERT structure are shown in Figure 4.4. Each encoder layer contains two sublayers: a position-wise feed forward network and a multiheaded self-attention network. BERT may be used to extract models that employ fine-tuning methods. When a model understands the language but is unable to use it to solve an issue, fine tuning procedures are used. We employ BERT in sentiment analysis since it enables us to comprehend how consumers feel about various goods, such as movies and other media, and how they rate them. Data collection is the initial stage in sentiment analysis. the data is gathered from various social media platforms, like Facebook, Twitter, etc. The data is labelled, and then it is sent to pre-Processing, where it is checked for missing values, noise, spelling errors, feature extraction, and dimension checks. A 7:3 ratio was used to split the datasets into training and test Segments.

The model was then analyzed in BERT, where it has two variants: BERT-base and BERT-large, using supervised learning. We have 12 years and 12 heads in the BERT-base model and 24 years and 16 heads in the BERT-large model. As BERT is bi-directional, it transmits data from both the left and right sides of a token. The two deep network layers employed in the BERT model are CNN and LSTM. Convolutional neural network, or CNN in its full name, is used for text categorization and recognition. The covulation layer and the max pooling layer are the two layers of CNN. Long Short-Term Memory, or LSTM for short, is a sort of RNN used for language production and translation. LSTM and CNN are not directly used by BERT because it is a transformer-based model. It made use of a self-attention apparatus to consider each word's text in the incoming data.

Technologies used:

Apache Hadoop

Apache Hadoop is one of the best open vendor platforms which offers high performance safe scalable paid form of large information unit processing to use simple coding model Hadoop uses clusters for computer systems that offer cost effective solutions Without the need for organized data, store and manage huge numbers of structured, semi-structured, or unstructured records.

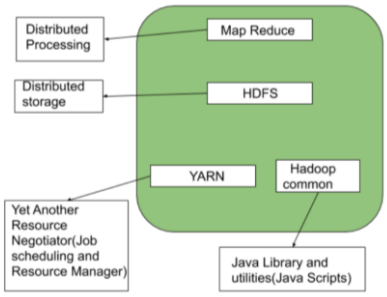

Figure 4.5: Components of Apachi Hadoop

Systems using different components of Hadoop works best for creating statistical sections that hold massive facts analytical initiative Hadoop helps site visitors with its own real-time operator in your data research business or other line of businesses. The components constituting Hadoop are portrayed in Figure 4.5. It is powering the advancement of statistical technology an interdisciplinary discipline which puts together device control information expert assessment and programming however the framework of Hadoop presents several challenging situations especially regarding the timing Hadoop can be very complicated and building maintaining and improving it requires great resources and know-how it is also time-consuming and inefficient due to the common parsing and notation used for computation.

Artificial Intelligence

The development of live intelligence processes by computing systems is referred to as artificial intelligence. AI comprises functions which include machine learning, natural language processing, speech/voice recognition, and expert computing systems. AI falls into two categories. Weak AI, frequently referred to as trained AI, is intent on doing a single job. Much of the AI that surrounds us is weak AI. The word "narrow" could be a more accurate description of this kind of AI because it isn't at all weak. It offers competent duties, just as Siri, Watson and Bing, and standalone automobiles. Strong AI is composed of ASI and AGI (Artificial General Intelligence)

(Artificial Superintelligence). Theoretically, machines will have intelligence identical to humans thanks to Artificial General Intelligence (AGI). Its capacity for learning, problem-solving, and making plans for the future will be weak. Artificial superintelligence (AI) will be more intellectual and powerful than the human mind. Strong AI currently has no legitimate applications, but this does not hinder researchers from exploring this field of study.

Machine Learning

Machine learning is a concept that allows computer systems to learn without any human beings having to program it it collects the specified information on its own and solves the challenge presented to it plays an vital role within the growth and upward thrust of artificial intelligence and use of neural networks, all of which involve machine learnings, pattern recognition capabilities It is now divided into four categories. The first is supervised learning, where the datasets we utilized are already labelled and help train the algorithm, facilitating it to analyze input more accurately and anticipate answers more swiftly. The second category is unsupervised learning, it clusters and analyses datasets without labels they then use this cluster to find patterns in the datasets with no human help Third, in semi-supervised learning, a smaller set of labelled data is input further into model and the algorithms then utilize those to detect any correlations within the dataset. This is advantageous while there isn't always enough labelled information because even a tiny volume of information can still be used to teach the machine. Last but not least, reinforcement learning, in which Through trial and error, the algorithm picks up new capabilities as it goes. Input information about whether a result was successful or failed is sent to the system Using a genuine text content dataset, Bert is a deep bidirectional unsupervised language depiction.

Semantic Analysis

Semantic analysis is the process of comprehending natural language (text) by sifting through unstructured data to extract relevant information like context, emotions, and attitudes. It makes it possible for computers and other systems to understand, analyze, and deduce meaning from sentences, blocks of text, report documents, registers, files, or anything similar textual content. Semantic analysis seeks to understand the organization of words, phrases, and clauses in sentence fragments in order to figure out the interconnections between individual pieces in a certain context.

Figure 4.6: Process of Semantic Analysis

SA natural language processing (NLP) do this job competently. The different processes that goes into SA is shown in Figure 4.6. It is also a crucial element of a number of machine learning products that are now on the market, including text analysis software, chatbots, and search engines. Lexical semantics, the initial design of semantic analysis is also often made reference to as the dictionary definitions and implications of certain terms. Then, in order to fully understand the context, the association between words in a sentence is examined. Vector space model uses Term Frequency-Inverse Document Frequency (TF-IDF) to evaluate the importance of words in a document. TF-IDF is treated as an important metric in text analysis domain.

Term Frequency (TF) measures the number of times a word occurs in the document.

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

where, $f_{t,d}$ is the occurrence of a term in a document divided by total numbers of terms in the document.

Inverse Document Frequency (IDF) evaluates the importance of a term appearing in the document.

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

where, N is the total number of documents in a corpus divided by those many numbers of document where the specific term appears.

Thus, TF-IDF is calculated by,

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

Sentiment Analysis

To assist businesses, monitor branding and product manufacturing sentiment in consumer feedback and understand their demands, sentiment analysis is a natural language processing technique that allows us to evaluate whether data is positive, negative, or neutral. The automatic monitoring and analysis of sentiment in all sorts of data via sentiment analysis is increasingly becoming a critical activity. Businesses may analyze what satisfies and irritates their consumers through comprehending client feedback, such as comments expressed in survey replies and discussions on social media platforms. People today express their emotions and thoughts more freely than before. It may be used as a stand-in for measuring customer satisfaction so that companies can adjust their products and services to

meet the needs of their target markets. it is a technique that makes use of text processing and natural language to try to understand the thoughts of a person. Businesses that launch new products would really like to understand consumer perception by gathering written feedback from customers within a matter of days after the product purchase. Sentiment analysis is a crucial tool for assessing out how people feel about a product and for coming up with practical ways to raise the quality in terms of goods and services offered. This can be supported by taking a simple example as shown in Table 4.2.

Table 4.2: Sentiment classification of reviews

TEXT	TAG
Best hotel. Beautiful location, good food, clean and hygienic rooms.	Positive
Beautiful ambience and location but service little slow and small rooms	Neutral
Pathetic service and overpriced hotel	Negative

Cloud Computing

Cloud computing is the delivery of computing services such servers, data storage, databases, networking, software, analytics, and intelligence over the internet in order to deliver flexible resources, rapid innovations, and economics of scale ("cloud"). Or to put it another way, businesses can rent accessibility to another party's infrastructure, such as storage, processing servers, and databases, from a cloud computing service provider and only pay for the resources they actually use. This is a substitute to building their own data centers.

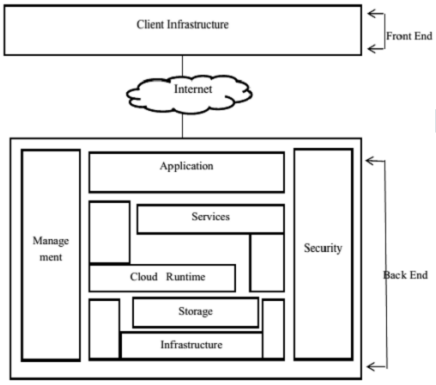

Figure 4.7: Architecture of Cloud Computing

Cloud computing is the delivery of computing services such servers, data storage, databases, networking, software, analytics, and intelligence over the internet in order to deliver flexible resources, rapid innovations, and economics of scale ("cloud"). Or to put it another way, businesses can rent accessibility to another party's infrastructure, such as storage, processing servers, and databases, from a cloud computing service provider and only pay for the resources they actually use. This is a substitute to building their own data centers.

To follow and understand how cloud computing works, let's divide it into its front end and back end, just how the Figure 4.7 shows. The front end is the client's computer or computer network. The front end is the customer's computer network. The back end of the cloud is formed of several computers, servers, and data storage systems. They are connected by way of a network, most often the Internet. The front end of a computer pertains to the client or user interface. The back end of the system is "the cloud" part.

V. CONCLUSION

This analysis of our project will help determine the user's interest in each tourist attractions. This will help increase the profits of tourist industry by being able to market the right aspects of each attraction to a specified demographic. It also helps increase SEO scores by utilizing the most used words as analyzed by the project. This is a simple yet efficient way to understand the user's preferences and provide suggestions based on the same analysis.

This study's findings provide an answer to a perennial query in the fields of tourism. What are the primary attractions that Travellers value and care about while visiting a certain location? This study seems to have some ramifications for restaurateurs and top management of destination marketers because it empirically shows the magnitude of the affiliations between the psychological thoughts expressed by visitors in the appraisal dialect and the electronic word-of-mouth (eWOM) of the holiday destination. Hence, destination management organizations (DMO) managers should concentrate on enhancing destination services, paying special attention to important factors that might lead to unfavorable WOM. The results of this study will help DMO managers identify destinations' weaknesses more precisely and boost their competitiveness more successfully. By studying and comparing all the types of ways we can analyze the customer's reviews, we can identify that BERT model works most efficiently and has accuracy more than KNN, SVM and Naves Bayes. Also, BERT model can be deployed over cloud easily, which is useful when there are huge sets of data to be processed.

VI. REFERENCES

[1] Hossein Hassani, Christina Beneki , Stephan Unger, Maedeh Taj Mazinani and Mohammad Reza Yeganegi, "Text Mining in Big Data Analytics", 2020.

[2] Ravi Vatrapu, Raghava Rao Mukkamala, Abid Hussain and Benjamin Flesch, "Social Set Analysis: A Set Theoretical Approach to Big Data Analytics", 2015, doi: 10.1109/ACCESS.2016.2559584.

[3] Muhammad Inaam Ul Haq, Qianmu Li and Shoaib Hassan, "Text Mining Techniques to Capture Facts for Cloud Computing Adoption and Big Data Processing", 2019, doi: 10.1109/ACCESS.2019.2950045

[4] Alexandra L'Heureux, Katarina Grolinger, Hany F. ElYamany,and Miriam A. M. Capretz, "Machine Learning with Big Data: Challenges and Approaches", 2020, doi: 10.1109/ACCESS.2017.2696365

[5] Jingyi Zhang, Tong Zhao, and Pingsheng Zhu, "Analysis Method of Motion Information Driven by Medical Big Data", 2019, doi: 10.1109/ACCESS.2019.2956803.

[6] Mingchen Feng, Jiangbin Zheng, Jinchang Ren, Amir Hussain, Xiuxiu Li, Yue Xi, and Qiaoyuan Liu, "Big Data Analytics and Mining for Effective Visualization and TrendsForecasting of Crime Data", 2019, doi: 10.1109/ACCESS.2019.2930410.

[7] Xianghua Fu, Jingying Yang, Janqiang Li, Min Fang, and Huihui Wang, "Lexicon-enhanced LSTM with Attention for General Sentiment Analysis", 2018, doi: 10.1109/ACCESS.2018.2878425.

[8] Zhi Li, Rui Li, and Guanghao Jin, "Sentiment Analysis of Danmaku Videos Based on Naïve Bayes and Sentiment Dictionary", 2020, doi: 10.1109/ACCESS.2020.2986582

[9] Yassin S. Mehanna, and Massudi Bin Mahmuddin, "A Semantic Conceptualization Using Tagged Bag-of-Concepts for Sentiment Analysis", 2021, doi:10.1109/ACCESS.2021.3107237.

[10] Wenjuan Luo, Fuzhen Zhuang, Xiaohu Cheng, Qing He, and Zhongzhi Shi, "Ratable Aspects over Sentiments: Predicting Ratings for Unrated Reviews", 2014.

[11] Staphord Bengesi , Timothy Oladunni , Ruth Olusegun and Halima Audu, "A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion from Twitter Tweets", 2023, doi: 10.1109/ACCESS.2023.3242290

[12] Rachmawan Adi Laksono , Kelly Rossa Sungkono and Riyanarto Sarno, "Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naive Bayes", 2019.

[13] Ren Cai, Bin Qin, Yangken Chen, Liang Zhang, Ruijiang Yang, Shiwei Chen and Wei Wang, "Sentiment Analysis About Investors and Consumers in Energy Market Based on BERT-BiLSTM", 2020, doi: 10.1109/ACCESS.2020.3024750.

[14] Yiren Chen, Xiaoyu Kou, Jiangang Bai, and Yunhai Tong, "Improving BERT With Self-Supervised Attention", 2021, doi: 10.1109/ACCESS.2021.3122273.

[15] Yan Cheng, Huan Sun, Haomei Chen, Meng Li, Yingying Cai, Zhuang Cai, and Jing Huang, "Sentiment Analysis Using Multi-Head Attention Capsules with Multi-Channel CNN and Bidirectional GRU", 2021, doi:10.1109/ACCESS.2021.3073988.

[16] Yuanzhao Gao, Xingyuan Chen and Xuehui Du, "A Big Data Provenance Model for Data Security Supervision Based on PROV-DM Model", 2020, doi: 10.1109/ACCESS.2020.2975820

[17] Keting yin, Shan Wang, Gang Wang, Zhengong Cai and Yixi Chen, "Optimizing deployment of VMs in cloud computing environment", 2013, doi: 10.1109/ICCSNT.2013.6967208

[18] Ashish Kumar Mishra, Abhishek Kesarwani and Dharmendra K Yadav, "Short Term Price Prediction for Preemptible VM Instances in Cloud Computing", 2019, doi: 10.1109/I2CT45611.2019.9033677

[19] Miao Chu, Yi Chen, Lin Yang and Junfang Wang, "Language interpretation in travel guidance platform: Text mining and sentiment analysis of TripAdvisor reviews", 2022, doi: 10.3389/fpsyg.2022.102994

[20] Les Servi and Sara Beth Elson, "A Mathematical Approach to Gauging Influence by Identifying Shifts in the Emotions of Social Media Users", 2014.

[21] Koyel Chakraborty, Siddhartha Bhattacharyya and Rajib Bag, "A Survey of Sentiment Analysis from Social Media Data", 2020.

[22] Mikel Zorrilla, Julián Flórez, Alberto Lafuente, Angel Martin, Jon Montalbán, Igor G. Olaizola and Iñigo Tamayo, "SaW: Video Analysis in Social Media with Webbased Mobile Grid Computing", 2017.

[23] Kwok Leung Tsui,Yang Zhao and Dong Wang, "Big Data Opportunities: System Health Monitoring and Management", 2019.

[24] Chinho Lina and Meichun Linb, "Application of Big Data in a MultiCategory Product-Service System for Global Logistics Supports", 2019.

[25] Dezhong Yao, Chen Yu, Laurence T. Yang and Hai Jin, "Using Crowdsourcing to Provide QoS for Mobile Cloud Computing", 2015.

[26] Vallikannu Ramanathan and T.Meyyappan, "Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism", 2019.

[27] Amit Kumar Goel and Kalpana Batra, "A Deep Learning Classification Approach for Short Messages Sentiment Analysis", 2020.

[28] Weiwei Lin, Ziming Wu, Longxin Lin, Angzhan Wen and Jin Li , "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis", 2017, doi: 10.1109/ACCESS.2017.2738069.

[29] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu, " Target-Dependent Sentiment Classification With BERT", 2019.

[30] Michele De Donno, Koen Tange, and Nicola Dragoni, "Foundations and Evolution of Modern Computing Paradigms: Cloud, IoT, Edge, and Fog", 2019.