World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:17, No:11, 2023

Web Data Scraping Technology Using Term Frequency Inverse Document Frequency to Enhance the Big Data Quality on Sentiment Analysis

Sangita Pokhrel, Nalinda Somasiri, Rebecca Jeyavadhanam, Swathi Ganesan

Abstract**—**Tourism is a booming industry with huge future potential for global wealth and employment. There are countless data generated over social media sites every day, creating numerous opportunities to bring more insights to decision-makers. The integration of big data technology into the tourism industry will allow companies to conclude where their customers have been and what they like. This information can then be used by businesses, such as those in charge of managing visitor centres or hotels, etc., and the tourist can get a clear idea of places before visiting. The technical perspective of natural language is processed by analysing the sentiment features of online reviews from tourists, and we then supply an enhanced long short-term memory (LSTM) framework for sentiment feature extraction of travel reviews. We have constructed a web review database using a crawler and web scraping technique for experimental validation to evaluate the effectiveness of our methodology. The text form of sentences was first classified through VADER and RoBERTa model to get the polarity of the reviews. In this paper, we have conducted study methods for feature extraction, such as Count Vectorization and Term Frequency – Inverse Document Frequency (TFIDF) Vectorization and implemented Convolutional Neural Network (CNN) classifier algorithm for the sentiment analysis to decide if the tourist's attitude towards the destinations is positive, negative, or simply neutral based on the review text that they posted online. The results demonstrated that from the CNN algorithm, after pre-processing and cleaning the dataset, we received an accuracy of 96.12% for the positive and negative sentiment analysis.

Keywords**—**Counter vectorization, Convolutional Neural Network, Crawler, data technology, Long Short-Term Memory, LSTM, Web Scraping, sentiment analysis.

I. Introduction

MOST people use social media to share their thoughts and opinions on services, goods, and places. These opinions are based on the feelings, reflection of the thoughts, and rating connected with a reference. Millions of people travel the world for work, recreation, sightseeing, or other purposes to meet their needs. Visitors will now have the chance from various sources of information that produce their content and make their components available easily. Internet sharing content of tourists has extremely prestigious tourism-related data. Currently, consumer opinions about hotels, restaurants, and places are critical components of the tourist industry which helps to make choices before they made their decision of booking plane tickets experience while visiting new places based on the sentiment polarity and will check the accuracy by using CNN algorithm.

The remainder of this paper is structured as follows: Section II illustrates how sentimental analysis is now in use and has been used to nourish the tourism industry. The background of previous similar papers on the system will be described in Section III, followed by the system's primary steps, while all the data collection, scraping, converting into sentiment polarity and CNN algorithm model are presented in Section IV. The proposed method about the Sentiment Analysis (SA) framework and algorithm development is discussed in Section V. The results with evaluation of findings, discussions and the conclusion are in Section VI and VII, respectively.

II. Background

According to the International Labour Organization 2019 [1], it is said that there is 1 job created in every 11 jobs based on tourism; so, it is no surprise why this sector plays such a significant role when increasing the world economy which currently stands at around $80 trillion dollars per year. It is estimated that by 2050, there will be more than 9 billion people living in countries around the world. This means that every year millions of people travel to visit places they have never seen before and experience new things. Travel and tourism contributed 8 trillion US dollars to the global GDP in 2019 [1]. In terms of employment, the World Travel & Tourism Council (WTTC) accounted that the tourism [1] sector supported 334 million jobs worldwide in 2019. For every 1 newly created job, nearly 1.5 added jobs were created through the indirect or induced effect on tourism-related economic activity [1].

Variety of data types are increasing over social media platforms (including Google Maps, Twitter, Instagram, Facebook, Flickr, etc.). So that the big data analytics is used to describe the process of defining, collecting, storing, retrieving, and analysing large data sets to understand their contents, and feelings and use them for decision-making [2].

The tracking of visitor behaviour is a significant challenge for Travel Destination (TD) management. TD managers need to be aware of the specifics of tourist hotspots, what draws visitors to each spot, how visitors themselves interpret their experiences, and how they intend to behave when they travel in the future. Most present approaches, in general, are unable to deal with these problems in a decision-centric, integrated, and

Sangita Pokhrel, MSc, Nalinda Somasiri, Head of the Programme, Rebecca Jeyavadhanam, Lecturer, Swathi Ganesan, Lecturer, Department of Computer Science, York St John University, London, United Kingdom (e-mail: sangita.pokhrel@yorksj.ac.uk, n.somasiri@yorksj.ac.uk, r.balasundaram@yorksj.ac.uk, s.ganesan@yorksj.ac.uk).

proficient manner. The majority of the currently used techniques for studying social media data are more concerned with providing answers to predetermined questions from their research than with gaining a broad understanding of how tourists move through their surroundings and what interests and experiences they have traveling, they do some research about the place where they will come across several websites, lengthy descriptive articles, records, etc. which is a challenging thing to understand such lengthy facts [3]. To make it easier we are doing the SA to make a proper update on their perspectives.

SA is the research that focuses on people's attitudes, perceptions, feelings, and views on things including people, organizations, services, events, subjects, places, and so on. The purpose of this paper is to determine if a particular review is subjective by representing the tourist's ideas or objectives by conveying their thoughts about the places, they visit [4]. Automated technologies like text mining are needed since they cannot be analysed manually which is the process of analysing the data from unstructured sources [5]. With the advent of Natural Language Processing (NLP), computer systems can easily read natural languages like English. SA makes it possible to understand the emotions and extract the hidden sentiments, attitudes, and expressions present in the textual content with the use of text mining and NLP techniques [6]. This paper develops an SA model and offers an overview of SA, text mining, and NLP in the tourist industry. Apart from these, there are many intelligent classification techniques such as Support Vector Machines (SVM), Bayesian classification, decision trees, and KNN algorithms and alternative text classification methods like text classification using neural networks. By building the semantic features and binary models for SA of comment data, it is needed to classify the text according to its emotional polarity and we are analysing all those things using CNN [7].

III. Literature Survey

Since the use of SA in the tourist industry has grown in popularity. Most of the research investigates sentiment categorization in tourism using machine learning techniques. The study of sentiment is a significant text-mining research topic where the analysis has been a hot topic in recent years. These papers cover a range of methods including objective and subjective phrases. Since the inspiration of our topic is based on previous works which we will evaluate first, the growth of regulatory frameworks and technological advancements have a direct impact on the tourist and hospitality industries. Automated SA, which has been used to extract thoughts from the public and a variety of other sources for varied uses in tourism, advertising, and hospitality, compares their effectiveness to that of human evaluators to estimate the compatibility of diverse types of automated classifiers [8].

In [9], the authors offered a method for SA of the movie reviews to evaluate the polarity of movies by collecting the data from the IMDB website. A special machine learning approach was used to break the number of separate phrases in which the sentence's overall meaning was determined by the synonyms for each word in the Tamil language. The sentences were grouped into several pre-established clusters, and they were categorized as similar sentences and belonged to the same cluster only if those sentences have the same meaning. And at last, the text elements were retrieved including the bag-of-words model, defining adjectives and adverbs, controlling negations, enforcing word frequency bounds, and using Word Net synonyms information. Reference [10] demonstrated the variety of NLP tasks benefits from the usage of language vector representations. In this research, the performance of word vector representations for SA is issued in three sub-tasks on prediction, word recognition, and extraction. To compute the numerous vector-based attributes, the methodological studies showed how successful they were in the analysis of APP reviews by obtaining F1 86.77%, recall 85.20%, and accuracy 86.35% using the straightforward vector-based features. In the research paper [11], we can conclude that the evaluation was utilized by a web crawler scraping methods based on R statistical software suite. Based on the concealed emotions of eight distinct categories, the tourist emotion distinction was helpful for the various restaurant categories. Text mining models for the decision support tool to analyse the intriguing remarks made by the visitors online that are broken into the stages: retrieving the documents on a particular topic, building a database, and removing stop words, stemming, clustering documents matrix, and at the result identifying the associations with a build word cloud.

Long Short-Term Memory (LSTM) approach is considered a moderately high classification approach that falls in the range from 0.73 to 0.85 on the AUC value. The refined dataset is tailored using an LSTM model for high-accuracy achievement and accurate prediction. According the research paper on [13], we can get an idea about managing the unbalanced dataset by the sampling techniques like under-sampling and over-sampling. In addition to categorizing for SA, this study evaluates the model's effectiveness by determining the recall, macro average, accuracy, F1-Score, and weight average values for each categorization of a beach. A CNN is a deep learning system that can take in an input picture, rank distinct parts of the image according to relevance, and distinguish between them [14]. The study develops an emotion classification model based on a CNN to categorize emotions. The model's precision was examined on the sentiment outperforms with the accuracy of 91.6%, demonstrating the model's usefulness. Comparatively speaking [15], the input picture was taken using the deep learning system CNN, ranked distinct parts of the image according to their relevance, and distinguish between them. A ConvNet's architecture was influenced by how the Visual Cortex is organized like the connection network of neurons in the human brain. ConvNet can learn the filters and properties in the constrained area of the visual field, known as the Receptive Field, where the fields are grouped and encompass the full image. To determine the optimum method for extracting data from internet sources, [16] described the data extraction using wrapper techniques. They did a single web page data extraction performance with several models, such as document object model, wrapper using hybrid JSON, and extraction of an image using Document Object Model (DOM), to determine the effectiveness of the suggested models. One of the important

aspects of this extraction procedure was that the execution time of bulk images with the filenames was reduced and it can extend up to many deep web pages in further work. The hospitality industry now has more chances for study since customers are increasingly researching and buying travel and hospitality services online. Despite numerous attempts to employ analytical models, experiments, or survey data to better understand the internet travel business, these studies frequently fall short of capturing the entire complexity of the sector. Another similar research [17] provides the tools and techniques needed to supplement the data sources easily and accessible information that travellers use to plan their trips. This paper offered instructions along with the Python code on how to gather/scrape data from any websites that are available online. Then the Python library Selenium was introduced for dealing with interactive, JavaScript-heavy pages like TripAdvisor, Airbnb, and Expedia and from other platforms too [18].

IV. Experimental Study

A. Web Scraping and Data Extraction

At present, one of the key issues while collecting the data is web traffic. For various purposes and significant advantages, people are more concerned about the method of extracting data from web pages. Since the website structures vary from one another and are semi-structured, manual selection of the data takes a lot of time. To eliminate this problem, the term 'Wrapper' comes as a useful time that is used for web data extraction (WDE). While dealing with useful extraction, post-processing may be necessary when using online data extraction to gather the information quickly for users and with great accuracy and recall.

DOM, visual segmentations, and other approaches have been used in several research to try and extract structured information from web pages [19]. Today, information is extracted using an extractor since the online pages have oceans of data that make finding information a difficult effort. The practice of extracting information from web documents in HTML format which may or may not be structured using tools and wrappers is known as WDE [27].

B. Methods of Web Scraping

As the World Wide Web developed, so did the techniques for web scraping. Not all the techniques were accessible at first. Since these are now the most popular methods, DOM, HTML parsing, XPath, and the APIs, it is used in applications for web mining, contact scraping, data mining, monitoring, and web indexing, and helps to gather real estate listings, weather data, monitor online presence and reputation, web mashup, and data integration. Moreover, few websites employ the techniques to stop web scraping by identifying and preventing bots from crawling their pages [20].

Manual Scraping

It is carried out manually where all we need to do is retrieve the data and keep it in a spreadsheet one by one. This is only suitable when the dataset we are expecting is small enough to copy and paste.

HTML Parsing

Transforming a program into an internal format where a runtime environment executes is known as parsing. The information on websites is not always available in convenient file types like .csv or .json files. The server responds to a request from the user by producing an HTML page. The offered web page's repeated parts can be seen by analysing the HTML structure and the output in the browser is what matters the most. Each page is a reoccurring pattern of source data which can be saved in the spreadsheet using a programming language script or web scraping software. HTML parsing is composed of tokenization and tree building. There are start and end tags, attribute names, and values, which are all included in HTML tokens. If the document is properly formatted, parsing it is simple and quick. Tokenized input is parsed into the document, creating the document tree.

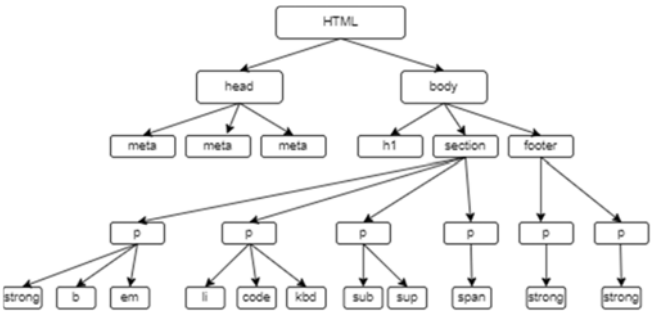

Fig. 1 HTML Structure

Document Object Model Parsing

DOM parser is designed to operate .xml data as an object graph tree structure in memory. First, the parser goes through the input XML file and constructs DOM objects that corresponded through the nodes. These DOM items are connected to one another in a manner like a tree. When the parser is finished, we receive this DOM object structure through which we can go back and forth to change, delete, or edit it. JavaScript and Cascading Stylesheets (CSS) make extensive use of the DOM [21]. Containers with individual DOM addresses are seen in Fig. 2. These are used in web scraping to make it simpler to navigate website content.

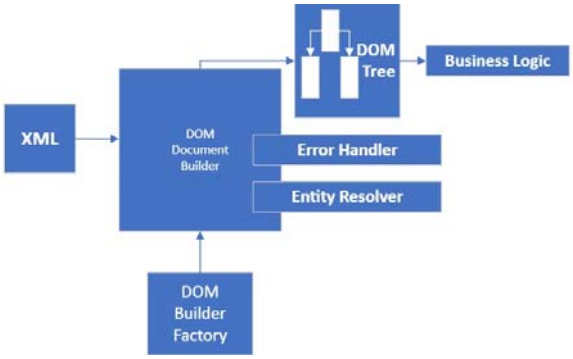

Fig. 2 DOM Parsing Selector

XPath

XPath is used by path expressions to choose specific XML

nodes or node sets. It offers a similar addressing option to that of the DOM in the HTML format as well. XPath defines a more finely organized webpage with a better ability to address webpage parts. To utilize it to parse a document, packages relevant to XML should be imported followed by document builder. The stream or file is converted to a document to make the path expression and object. The list of nodes is produced by compiling () and evaluating () in the repeated format [22].

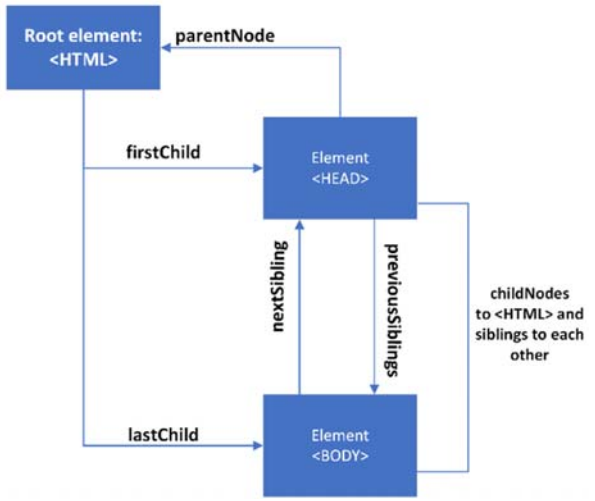

Fig. 3 XPath Navigation

API

Machine-readable interfaces are another term for application programming interfaces (APIs). Most of the APIs that are available are listed, explained, and provide pertinent connections to the sources in the directory. An API Endpoint receives a response from the server in response to a typical HTTP request. Each API has its own choices and specifications. The request can specify the format of the response. However, it is more of a feature that website and software owners can decide to use than a data extraction tool. APIs serve as a middleman, enabling communication and information sharing between websites and software. Nowadays, many websites that deal with enormous volumes of data, such as Facebook, YouTube, Twitter, and even Wikipedia, have a dedicated API. While an API is organized in its data extraction, a web scraper is a tool that lets you search and scrape the furthest reaches of a website for data [23].

The custom-made communication protocol used by a website or app might be compared to an API. It must first talk in its language and be accepted by specific laws to interact with it. We do not have to be concerned about using a proxy server or having our IP address blocked because they are an official tool provided by the website [24]. Additionally, APIs only offer access to the data the owner wishes to share, so we do not have to worry about breaking any moral boundaries and taking information that we were not supposed to do.

C. System Architecture

At first, we will crawl the review data from the web and the feature words are to be extracted in preprocessing stage. Since the multiple feature terms might relate to the same aspect, we will determine if a phrase refers to an aspect after classifying it into several algorithms. Later, we use POS to extract feature words based on the word structure. We will employ emotional words to quantify the sentiment strength of features: whenever a feature is modified by positive emotion, we remember to assign it a sentiment strength value of +1, and whenever a feature is modified by negative emotion, we remember to assign it a sentiment strength value of -1. The emotional intensity levels for each quality are then added together. The findings of our study highlight the tourist sentiments and will draw a conclusion.

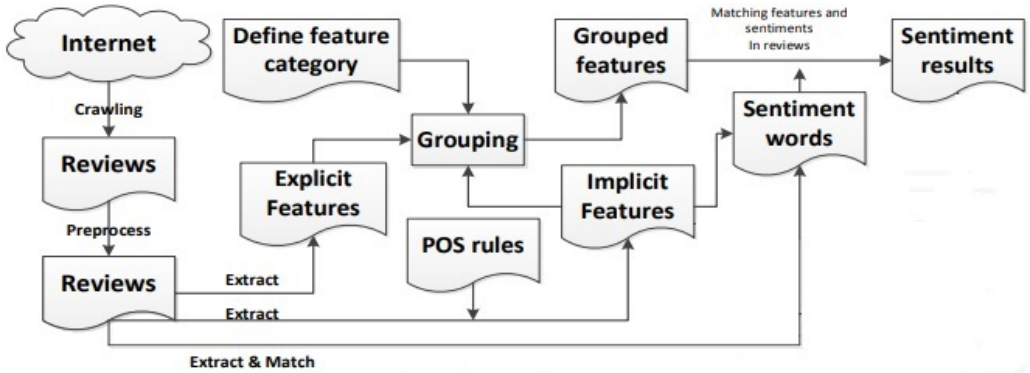

Fig. 4 System architecture in SA

SA may be broken down into two study levels for words and phrases, sentences, and chapters depending on how finely the text is processed. The characteristic or purpose of the review is generally an aspect, often known as a feature. For instance, the features of some areas include restaurants, roads, and temples which impact the services [25].

V. Proposed Method

In this proposed work, top ten tourist destinations were considered from Nepal. In visiting the website, our web browser sends a request to a web server, this request is known as GET request. After that, the server returns files that instruct our browser on how to display the page for us. Typically, these files

include the HTML for page. Our web browser sends a request to a web server whenever we access a web page. Since we are downloading files from the server, this request is known as a GET request. After that, the server returns files that instruct our browser on how to display the page for us. Typically, these files will include HTML for the main content, CSS for styling web pages and improve their appearance, Java Script to make webpages more interactive and images [26].

Procedure for scraping the data:

1) Send an HTTP request to the website's URL that you want to visit. When a request is made, the server responds by returning the webpage's HTML.
2) Parse the data after accessing the HTML content. Almost all the html data are nested so that string processing itself cannot be used simply to extract data. There are many HTML parser libraries, and among them html5lin is the advanced and most used one.
3) Use the Python library BeautifulSoup to search and extract the data from the parse tree.

A. SA Framework

A subfield of artificial intelligence known as machine learning focuses on designing and creating algorithms that enable computers to evolve behaviours based on empirical data, such as that gathered from databases or sensor data. Examples (data) can be used by a learner to identify key features of an unidentified underlying probability distribution. The goal of machine learning research is to develop software that can automatically identify intricate patterns and draw informed conclusions from data. This goal is challenging since the set of all potential behaviours for all potential inputs is too large to be covered by the training data set of observed examples so we will split the data into training and test datasets.

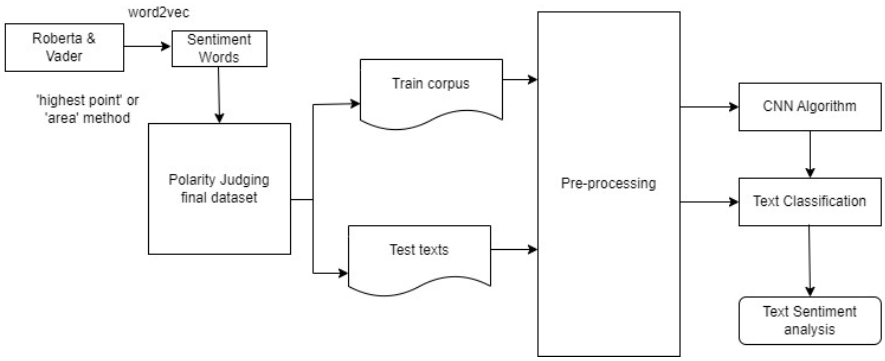

Fig. 5 Overall Architecture of the Proposed System

B. Pre-Processing of Data

Pre-processing the data is an important step since it affects how well the subsequent processes work. The sentences must be contextually corrected as needed. For the measures to decrease uncertainty in feature extraction, the data should be made more machine readable.

The processes for pre-processing data are listed below:

i. Duplicate reviews are removed from the dataset to ensure data quality.
ii. Since we are utilizing the case sensitive analysis, we should convert all the words into upper cases to lower case for a good analysis.
iii. Stop words that have no significance on the text of the sentences are eliminated (such as and, or, yet, etc.).
iv. Unnecessary columns like usernames, locations, and URLs are either removed or replaced with generic tags.
v. Stemming is applied to replace words with their origins and to group various word forms with similar meanings.
vi. Digits and special characters should be removed since they lack emotional meaning. Sometimes they are combined with words, therefore getting rid of them might assist connect two words that would otherwise be viewed as distinct.
vii. A dictionary is employed to filter out unwanted words and punctuation, and spelling is improved. Additionally, a dictionary is generated for important words and emotions.
viii. Part of Speech (POS) tagging: It assigns the word by classifying into specific category such as noun, adverb, adjective etc. These taggers are really useful for explicating the feature extraction.

C. VADER and RoBERTa SA

As we know, a computer understands only binary language i.e., 0 and 1. With the help of NLP, we will convert the human language to the numbers in terms of polarity. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a vocabulary and rule-based SA tool used to extract the sentiments expressed in social media. It is an open-source model to analyse the text sentiment to both polarity (positive/negative) and strength of emotion by returning a sentiment score in the range -1 to 1, from most negative to most positive. The sentiment score is determined by adding the sentiment of all the words in the sentence that are listed in the VADER dictionary.

For applications like evaluating public opinion, customers experience, product and place review, VADER approach might not interpret the sarcasm and irony sentences. Grammatical and misspellings errors may cause this analysis not much accurate. Hence, there is another model named as RoBERTa model. RoBERTa model is an AI developed tool which is trained on more than 124 M tweets for self-supervised NLP. Since the Roberta model is more accurate, we have taken the result and

converted to the sentiment score of -1, 0 and 1 doing the if else condition based on the RoBERTa model. By now, we have got a sentiment of all the reviews. We are visualizing the sentiments of only positive, negative based on word clouds, accuracy, and precision by applying CNN algorithms.

D.Algorithm Implementation

CNN is a classification tool that we are using for a mixed comment to indicate the positivity and negativity of a sentence in our data categories. It has previously had enormous frame in computer vision and speech processing work which now have improved in outcomes in NLP as well in the jobs like phrase modelling, semantic parsing, and query search [25]. Using Keras and TensorFlow, we will train a recurrent neural network model for this model. It is a little bit simpler to interface with Google's TensorFlow deep learning platform when using the Python deep learning library Keras, which can be run on top of a variety of different platforms. TensorFlow is really a strong system that is utilized extensively in business and research as well as behind the scenes at Google.

E. Training Algorithm for CNN Model

- Model: CNN_03: Sequential
- E = Embedding (100000, 200)
- Weights = Embedding Matrix
- Input Length = 45
- Trainable = True
- Filters = 100
- Kernal Size = 2
- Strides = 1
- Dense = 256: Activation = relu
- Dense = 1: Activation = sigmoid
- Epochs = 5

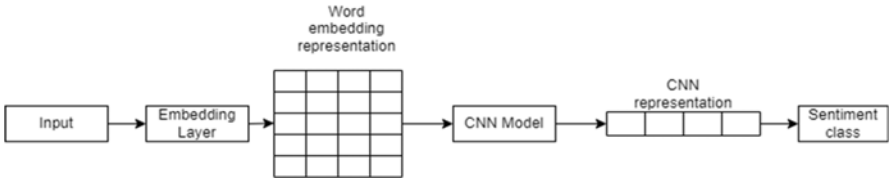

Fig. 6 Framework of CNN

The input data with the sentiment scores 1 and 0 are passed through the embedding layer to word embedding representation where the trained data are sequenced to the number of word count using tokenizer. Then the matrix generated in sequential validation index is trained by CNN model. After the CNN representation, the loss, accuracy, val_loss and val_accuracy can be generated for the sentiment class.

VI. Testing and Implementation

After gathering the data, we cleaned it by converting all text to lowercase, removing unnecessary elements such as punctuation, links, and text within square brackets. We improved the clarity of the text data by dividing it into individual words (tokenization), simplifying the words to their core form (stemming), and changing them to their basic form (lemmatization) to make it more consistent and easier to understand for the Roberta model while classifying each word as positive (1) or negative (0). Then we analysed the results using the CNN algorithm.

The CNN used the techniques of word embeddings to represent the distinct words with similar meanings in a comparable real-valued vector. They represent the significant advance that has resulted in excellent performance of neural network models on a variety of difficult NLP challenges. On the training set of sequential model of CNN, we got an accuracy of 96.01% with the validation accuracy of 98.15%.

VII. Results and Discussion

Following the pre-processing and the feature extraction procedures, we go on to training and testing the model's performance. In this section, we analyse the CNN algorithms of data to accomplish the SA tasks. We used Python programming language to define and analyse the SA in the reviews of tourist from top 10 destinations from Nepal and Jupyter notebook as a web-based interactive computing platform.

In terms of subjectivity, all the attractions investigated had a range of values ranging from 0 to 1, indicating that the evaluation provided by visitors is an opinion or personal viewpoint. Additionally, a wordlist is generated from the study to illustrate the frequency of each word. The result shows that only the 4% of the tourist among the collected 7 thousand data had the negative reviews. Fig. 7 shows the most common words used in the positive comments like best, beautiful, amazing, good, etc., which implies that most of the tourist who visited those places were happy and had enjoyed their visit.

TABLE I
Performance of the CNN Model

Epoch	Loss	Accuracy	Val_loss	Val_accuracy
1/5	0.6570	0.9522	0.6182	0.9815
2/5	0.5903	0.9601	0.5592	0.9815
3/5	0.5324	0.9601	0.4959	0.9815
4/5	0.4823	0.9601	0.4464	0.9815
5/5	0.4390	0.9601	0.4034	0.9815

Based on the findings, this study concludes that the tourists are pleased with the attraction as well as they have a nice experience during their visit to Nepal. Similarly, we can check the negative and neutral words based on the same word cloud technique. In this research, we employed web mining and Python programming tool to get insight on tourist attitude about tourism attractions of Nepal. Most of the data reveal that each tourist site CNN for epoch 3 of model training, we get the

following result in terms of loss, accuracy, validation loss and validation accuracy with the range of time per epoch.

In the case of CNN, we can see that the validation accuracy is constant for all the five epochs of 98.15% and the accuracy of 95.22% from the first epoch increased to the 96.01% to all other four epochs.

Fig. 7 Positive words for word cloud

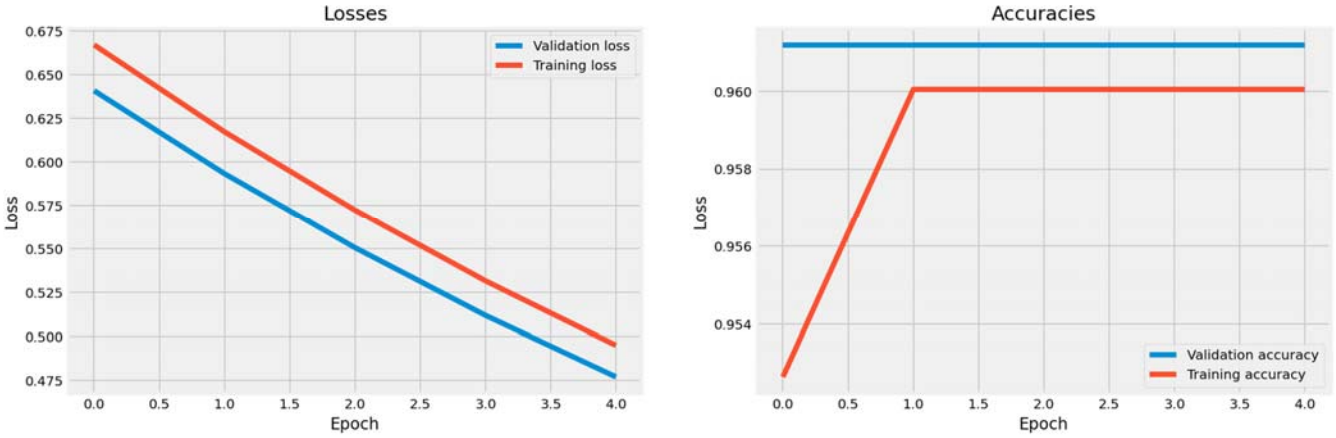

Fig. 8 Accuracy and losses plot

Fig. 8 shows the relationship in between the losses and accuracies of training and validation data from the model where we can compare the model's performance in a graphical format. While increasing every epoch, accuracy should go higher, and loss goes lower for the better prediction. The possible cases with the val_loss and val_acc are:

1. Val_loss starts increasing, val_acc starts decreasing that means model is not learning properly.
2. Val_loss starts increasing, val_acc also increases that might be a sign of overfitting.
3. Val_loss starts decreasing, val_acc starts increasing, which is correct and indicates that the model build is learning and working perfectly fine.

VIII. Conclusion

In the case of CNN, we have split data into two parts for every epoch i.e., training and validation data. The model is trained on training data and validation data are used for checking the accuracy of the model with the loss entity. There is an accuracy of 96.12% and loss of 59.32% for the loaded CNN model. Since we have tried the three sequential for keras.model, for all of them, the val_loss is decreasing and val_acc starts increasing and similarly loss and accuracy also

follows the same pattern. Thus, we can conclude that the CNN is learning and working accurately. We believe that the combined volume of study and a huge number of datasets can further inspire other academics to employ more deep learning models such as bidirectional recurrent neural networks to grasp the emotion in tourist reviews even more precisely with the much more data to concentrate in depth performance. This analysis can further be utilized by building a system to store the results in a database and make it visible in different graphs and techniques to understand by any simple person easily.

REFERENCES

[1] S. R. Department, "Total contribution of travel and tourism to gross domestic product (GDP) worldwide from 2006 to 2021," *Travel, Tourism & Hospitality,* no. 2022, 2022.

[2] M. H. A. Gandomi, " Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management,* no. 2022, pp. 137-144, 2015.

[3] G. F. S. T.March, " Design and natural science research on information technology," *Decision Support Systems,* no. 2022, pp. 251-256, 1995.

[4] D. R. R. D. Chingakham Nirma Devi, "Literature Review on Sentiment Analysis in Tourism," *Test Engineering and Management,* vol. 83, pp. 2466-2474, 2020.

[5] Renganathan, "Text mining in biomedical domain with emphasis on document clustering," *Healthcare Informatics Research,* vol. 3, no. 23, pp. 141-146, 2017.

[6] Q. C. C. S. E. S. P. Jiang, "Sentiment analysis of online destination image," *Current Issues in Tourism,* vol. 4, no. 26, pp. 1-22, 2021.

[7] A. M. a. I. M. Abubakar, "Impact of online WOM on destination rust and intention to travel: a medical tourism perspective," vol. 5, pp. 192-201, 2016.

[8] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, "CNN for situations understanding based on sentiment analysis of twitter data," *ResearchGate,* vol. 4, pp. 376-381, 2017.

[9] C. S. M. B, "An Approach of Sentiment Analysis for Movie Reviews," *International Conference on Communication, Computing and Internet of Thing,* 2022.

[10] X. L. F. D. X. L. M. W. Xian Fan, "Apply Word Vectors for Sentiment Analysis of APP Reviews," *The 2016 3rd International Conference on Systems and Informatics (ICSAI 2016),* 2016.

[11] A. U. Vinaitheerthan Renganathan, "Dubai Restaurants: A Sentiment Analysis," vol. 14, no. 2, 2021.

[12] E. S. P. W. Afina Ramadhani, "LSTM-based Deep Learning Architecture of Tourist Review in Tripadvisor," *Sixth International Conference on Informatics and Computing (ICIC),* 2021.

[13] Ali Aggaa, Ahmed Abbou, Moussa Labbadib, Yassine El HoumaImane, HammouOu Alia, "CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production," *Electric Power Systems Research,* vol. 208, 2022.

[14] T. Huang, "Research on Sentiment Classification of Tourist," *IEEE 3rd Eurasia Conference on IOT Communication and Engineering (ECICE),* 2021.

[15] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie & Laith Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data,* vol. 53, 2021.

[16] M. M. Ily Amalina Ahmad Sabri, "A deep web data extraction model for web mining: a review," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 23, pp. 519-528, 2021.

[17] Saram Han and Christopher K. Anderson, "Web Scraping for Hospitality Research: Overview Opportunities, and Implications," *Cornell Hospitality Quarterly,* 2021.

[18] A. Rao, "Convolutional Neural Network Tutorial (CNN) – Developing an Image Classifier in Python Using TensorFlow," Edureka, 15 09 2022. (Online). Available: https://www.edureka.co/blog/convolutional-neural-network/. (Accessed 11 2022).

[19] Z. Cai, J. Liu, L. Xu, C. Yin, J. Wang, "A Vision Recognition Based Method for Web Data Extraction," *Computer Science,* 2017.

[20] R. Mitchell, "Web Scraping with Python," O'Reilly Media, 2015.

[21] V. Draxl, "Web Scraping Data Extraction from websites," no. 2022, 2018.

[22] A. OT, "Web Scraping vs. API: What's the Best Way to Extract Data?," 2021. (Online). Available: https://www.makeuseof.com/web-scraping-vs-api/. (Accessed 03 09 2022).

[23] C. P. Colombage, "Comparing Deep Learning Architecture for Sentiment Assessment for Online Consumer Reviews," York St. John University – London Campus, Department of Computer Science, London, 2021.

[24] A. Sharma, "A guide to web scraping in Python using Beautiful Soup," 2021. (Online). Available: https://opensource.com/article/21/9/web-scraping-python-beautiful-soup. (Accessed 09 2022).

[25] A. R. V. R. C. A. R. D. A. K. M. a. S. K. Shalini K, "Sentiment Analysis of Indian Languages using Convolutional Neural Networks," *International Conference on Computer Communication and Informatics (ICCCI -2018),* no. 2022, 2018.

[26] Renganathan, "Text mining in biomedical domain with emphasis on document clustering," *Healthcare Informatics Research,* vol. 3, no. 2022, pp. 141-146, 2017.

[27] Chang, Chia-Hui and Shao-Chen Lui. "IEPAD: information extraction based on pattern discovery." *The Web Conference* (2001).