SENTIMENT ANALYSIS:

Sentiment analysis is a domain of natural language processing (NLP) plays a vital role in categorization of polarity into three outcomes such as positive, negative and neutral based on text reviews extracted from feedback given by consumers through online platforms like TripAdvisor and yelp datasets. sentiment analysis is utilized for various types of classification such as subjectivity of text data, emotion detection and justification of rating etc. Sentiment classification is currently become a key method for all kinds of business firms to know the effectiveness of services offered to end users and help to resurge in adverse conditions by changing their strategies according to current scenario by the integration of sentiment analysis.

The process of achieving sentiment classification involves mainly four steps a) import big data from internet sources b) data cleaning c) Text preprocessing d) model building with deep neural network algorithm using Kera’s and TensorFlow.

1. import big data from internet sources: Download TripAdvisor csv file from one of the online sources like Kaggle and read the csv file by deploying pandas’ package in PyCharm environment. set the path for pandas to read the csv text file.
2. Data cleaning: recycling of text data includes eliminating URL’s, html tags, punctuations, single character, multiple spaces.
3. Text preprocessing and normalizing labels: preprocessing of text requires feature selection and representation of bigrams and word embeddings. Normalizing label needs one hot encoder and label encoder to convert strings into integer.
4. Model building with deep neural network algorithm: firstly, construct deep neural network by adding word embedding layer as initial layer form input of word embeddings. Secondly, integrating word embedding layer input to long short-term memory (LSTM). finally, LSTM input is added into dense layer includes

Use pandas to read the csv file and remove null values in text data.

Bigrams: initially, Bigrams are derived from N-gram. N-gram is adjacent sequences of n-items like words, symbols and text from a given corpus text or document. Where “N” is a range to be selected for feature extraction of given data.in this project, consider N=2 which is union of two words in a text or sentence known as “bigrams”. For example, in the given array

[“the sun rises in the East”] in this array, the bigrams are given as “the sun”, “sun rises”, “rises in”, “in the”, “the East”.

N-grams are mainly used to distribute long sequences text data, sentences into small pair of words and to calculate occurrences of frequencies of input ‘n’ value from text corpus.

ONE HOT ENCODING:

One hot encoding is one of the preprocessing methods to encode categorical variables into binary values represented in 1’s and 0’s.in this process each input categorical string and integer are transformed into numerical values by creating new attributes for each category in a binary column matrix.one hot encoder denotes null value if the token s not presented in respected category

we saw how categorical variables such as the day of week can be one-hot encoded to numerical variables by creating a new feature for each variable.