REVIEW RATING USING TRIP ADVISOR

S. Ramyateja[1], A. Saivardhan Reddy[2], G. Varshini[3]

[1]Department of Electronics & Communication Engineering, SR Engineering College, Warangal, Telangana, India.

[2]Department of Electronics & Communication Engineering, SR Engineering College, Warangal, Telangana, India.

[3]Department of Computer Science & Engineering, SR Engineering College, Warangal, Telangana, India.

Abstract: Sentiment or opinion analysis employs natural language processing to extract a significant pattern of knowledge from a large amount of textual data. Sentiment analysis is a natural language processing tool that is useful for monitoring applications, as it can reveal public opinion about numerous issues without requiring satisfaction inquiries. The availability of a huge volume of reviews makes it troublesome for service executives to know the percentage of reviews that affect their services. Thus, developing a sentiment assessment technique concerning hotel reviews is essential, particularly in Indonesia. This research uses the Long- Short Term Memory (LSTM) and Word2Vec models. The integration of Word2Vec and LSTM variables used in this research are Word2Vec architecture, Word2Vec vector dimension, Word2Vec evaluation method, pooling technique, dropout value, and learning rate. On the basis of experimental research performed through 555500 review texts as a dataset, the best performance was obtained and had an accuracy of 85.96%. The parameter combinations for Word2Vec are Skip-gram as architecture, Hierarchical SoftMax as an evaluation method, and 300 as vector dimension. Whereas the parameter combinations for LSTM are a dropout value is 0.2, pooling type average pooling, and a learning rate is 0.001.

1 Introduction

Sentiment analysis (also referred to as subjectivity analysis or opinion mining or emotion artificial intelligence) is a natural language processing (NLP) technique that identifies important patterns of information and features from a large text corpus. It examines comments, opinions, emotions, beliefs, views, questions, preferences, attitudes, and requests communicated by the writer in a string of text. It extracts the writer's feelings in the form of subjectivity (objective and subjective), polarity (negative, positive, and neutral), and emotions (angry, happy, surprised, sad, jealous, and mixed). It analyzes thoughts, attitudes, views, opinions, beliefs, comments, requests, questions, and preferences expressed by an author based on emotion rather than a reason in the form of text towards entities like services, issues, individuals, products, events, topics, organizations, and their attributes. It finds the author's overall emotion for a text where text can be blog posts, product reviews, online forums, speech, database sources, social media data, and documents. It usually consists of three elements depending on the context:

1. Opinions or emotions: An opinion is also referred to as polarity, whereas emotions can be qualitative such as sad, joy, anger, surprise, disgust, or happiness, or quantitative such as rating a movie on a scale of one to ten
2. Subject: It refers to the subject of the discussion where one opinion can discuss more than one aspect of the same subject, for instance, the camera of the phone is great, but the battery life is disappointing.
3. Opinion holder: It refers to the author/person who expresses the opinion.

Natural Language Processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

The main objective of this study is to classify customer reviews into positive or negative sentiments, to measure the intensity of the sentiments generated by the customer, to analyze the association between customer reviews concerning different attractions in the city.

2 Literature review

[1] Sentiment Analysis with KNN Algorithm proposed by Hyunwoo Max Ch. A machine learning model that can predict if the person thinks positively or negatively about the movie based on the movie review data. To this end, the given data were first preprocessed to turn theminto a dataset suitable for training. Then, the machine learning model using the KNN algorithm was trained and the model was able to predict peoples' sentiments with 82.5% accuracy. Since the above program uses the KNN algorithm, it does not need to know how negative or positive the word.

[2] Sentiment Analysis Approach Based on N-gram and KNN Classifier proposed by Sumandeep kaur, Geetha. The proposed approach is a combination of feature extraction and classification techniques. The N-gram algorithm is applied for the feature extraction and KNN classifier is applied to classify input data into positive, negative, and neutral classes. To validate the proposed system, performance is analyzed in terms of precision, recall, and accuracy. The results of the experiment of the proposed system show that it performs well compared to the existing system which is based on an SVM classifier.

[3] Opinion analysis of travelers based on tourism site review using sentimental analysisintroduced by Siti Azza Amira, Mohammad Isa Irwan*. The support vector machine method combined with TF-IDF can solve problems in sentiment classification. This is evidenced by the ability of the TF-IDF method to give a weight value to a word and the ability of the Support vector machine method to provide labels in each review, which are positive reviews andnegative reviews. With this value of accuracy, it means the classifier used has worked well in classifying reviews.

[4]Opinion analysis of travelers based on tourism site review using sentimental analysisintroduced by Siti Azza Amira, Mohammad Isa Irwan*. The support vector machine method combined with TF-IDF can solve problems in sentiment classification. This is evidenced by the ability of the TF-IDF method to give a weight value to a word and the ability of the Support vector machine method to provide la-

bels in each review, which are positive reviews andnegative reviews. With this value of accuracy, it means the classifier used has worked well in classifying reviews.

[5] The influence of Trip Advisor application usage towards hotel occupancy rate in Solo proposed by D Sumarsono1,2,3*, B Sudardi1, Warto1, W Abdullah. Trip advisor also plays the role as a reference of the world tourism industry in raising the rating of the hotel. Trip advisor plays the role as a reference of the worlds tourism industry in raising the rating of the hotel.

3 Methodology

3.1 Problem Statement

The main objective of this study is to classify customer reviews into positive or negative sentiments, to measure the intensity of the sentiments generated by the customer, to analyze the association between customer reviews concerning different attractions in the city.

3.2 Data Insights

Address na	Additional	Review_Da	Average_S	Hotel_Nar	Reviewer_	Negative_	Review_Tc	Total_Nun	Positive_R	Review_Tc	Total_Nun	Reviewer_	Tags	days_since	lat	lng
s Gravesa	194	########	7.7	Hotel Aren	Ireland	No Negati	0	1403	No real cc	105	7	7.5	[' Leisure t	0 days	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	Italy	No Negati	0	1403	This hotel	59	6	9.2	[' Business	30 days	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	Italy	No Negati	0	1403	This hotel	82	26	10	[' Leisure t	31 days	52.36058	4.915968
s Gravesa	194	6/29/2017	7.7	Hotel Aren	Netherlan	No Negati	0	1403	Public are	33	4	7.1	[' Business	35 days	52.36058	4.915968
s Gravesa	194	3/22/2017	7.7	Hotel Aren	United Kir	No Negati	0	1403	The qualit	77	3	10	[' Leisure t	134 day	52.36058	4.915968
s Gravesa	194	3/16/2017	7.7	Hotel Aren	United Kir	No Negati	0	1403	Beautiful	49	4	10	[' Leisure t	140 day	52.36058	4.915968
s Gravesa	194	2/20/2017	7.7	Hotel Aren	United Kir	No Negati	0	1403	The hotel	76	2	10	[' Leisure t	164 day	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	Switzerlar	No Negati	0	1403	Basically e	84	16	9.6	[' Leisure t	175 day	52.36058	4.915968
s Gravesa	194	12/13/201	7.7	Hotel Aren	United Kir	No Negati	0	1403	The whole	56	1	10	[' Leisure t	233 day	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	United Kir	No Negati	0	1403	Hotel was	68	1	9.2	[' Leisure t	236 day	52.36058	4.915968
s Gravesa	194	9/27/2016		Hotel Aren	United Kir	No Negati	0	1403	We upgra	38	1	10	[' Leisure t	310 day	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	United Kir	No Negati	0	1403	Architectu	115	4	10	[' Leisure t	394 day	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	United Kir	No Negati	0	1403	Breakfast	34	1	9.6	[' Leisure t	397 day	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	United Kir	No Negati	0	1403	This hotel	78	1	8.8	[' Leisure t	397 day	52.36058	4.915968
s Gravesa	194	4/22/2016	7.7	Hotel Aren	United Kir	No Negati	0	1403	Beautiful	40	2	10	[' Leisure t	468 day	52.36058	4.915968
s Gravesa	194	3/17/2016	7.7	Hotel Aren	Ireland	No Negati	0	1403	Bar and re	33	1	6.7	[' Leisure t	504 day	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	United Kir	No Negati	0	1403	The staff	35	1	10	[' Leisure t	542 day	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	United Kir	No Negati	0	1403	The hotel	66	11	10	[' Leisure t	548 day	52.36058	4.915968
s Gravesa	194	1/23/2016	7.7	Hotel Aren	Ireland	No Negati	0	1403	Stayed in	78	1	10	[' Leisure t	558 day	52.36058	4.915968
s Gravesa	194	11/15/201	7.7	Hotel Aren	United Kir	No Negati	0	1403	Staff were	37	1	9.2	[' Leisure t	627 day	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	United Kir	No Negati	0	1403	Staff were	47	3	9.2	[' Leisure t	631 day	52.36058	4.915968
s Gravesa	194	########	7.7	Hotel Aren	Australia	No Negati	0	1403	Good loca	37	17	10	[' Leisure t	641 day	52.36058	4.915968
s Gravesa	194	10/29/201	7.7	Hotel Aren	United Kir	No Negati	0	1403	Hotel and	67	5	9.2	[' Leisure t	644 day	52.36058	4.915968
s Gravesa	194	10/22/201	7.7	Hotel Aren	Ireland	No Negati	0	1403	This was c	31	2	10	[' Leisure t	651 day	52.36058	4.915968
s Gravesa	194	10/17/201	7.7	Hotel Aren	Canada	No Negati	0	1403	It was a w	35	1	10	[' Leisure t	656 day	52.36058	4.915968
s Gravesa	194	9/29/2015	7.7	Hotel Aren	Spain	No Negati	0	1403	I loved the	67	1	9.6	[' Business	674 day	52.36058	4.915968

Fig 1. Dataset

1. We have collected the dataset from the internet Kaggle, it consists of address name, an additional number of scorings, review date, average score, hotel name, reviewer nationality, negative review, review total negative word count, a total number of reviews, positive review, review total negative word count, a total number of reviews reviewer has given, reviewer score, tags, days since the review, lat(longitude) and lng(longitude). The dataset has 17 columns along with 515739 rows.

2. Of this data, we have removed some rows and columns and made the number of rows finally399336 and the number of columns 4. The number of rows is decreased so that the processing of data takes less time compared with huge amounts of data and data can be accurately predicted.

3. The output feature predicts which trip is best for the customer

3.3 Data Preprocessing

Data pre-processing is essential while working on large datasets because algorithms could only be applied to vectorized text. Data pre-processing thereby aims at covering text in a vectorized simple form which means tokenizing. Tokenizing means dividing the text into units of words or sentences. Tokenizing is the fundamental step for stemming and lemmatization.

- **Tokenization**

Tokenization is the first step in any NLP pipeline. A tokenizer breaks unstructured data and natural language text into chunks of information that can be considered as discrete elements. The token occurrences in a document can be used directly as a vector representing that document. This immediately turns an unstructured string (text document) into a numerical data structure suitable for machine learning.

1. We have eliminated stop words from the dataset as they have no significance in deciding the meaning of the text. Stop words are the most common words in any language. By removing these words, we remove the low-level information from our text

in order to give more focus to the important information, and reduces training time. Stemming has been applied to correlate the words belonging to the same root. Stemming is the process ofreducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.

2. Then we would proceed with label encoding, the label encoding is to signify the categorical data for the semi-structured or unstructured data. Label encoding means giving the labels for the data in numerical. Next, stemming is done to produce morphologicalvariants of a root/base word.

3. Stemming programs are commonly referred to as stemming algorithms or stemmers. Thesealgorithms are used to give the domain vocabularies in domain analysis.

4. Neural networks require to have input of the same size. Therefore, sentence inputs are added with 0's after defining the max length and words are dropped and added accordingly

3.4 Methodology

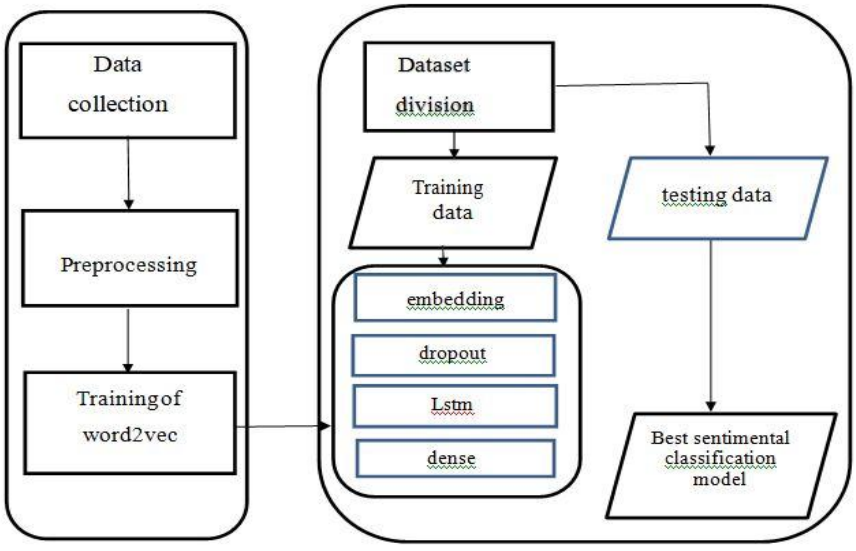

Fig 2. Block diagram

We collected the data set from Kaggle and uploaded that dataset into google colab with the help of google drive. The data thus obtained is preprocessed using some techniques like tokenization, stemming etc., then the data is trained using word2vec model followed by layers such as embedding, dropout, lstm and dense. Finally, the data is tested and the accuracy obtained is 86 percent.

➢ Word To Vector

Word2Vec model is used for Word representations in Vector Space which is founded by Tomas Mikolov and a group of research teams from Google in 2013. It is a neural network model that attempts to explain word embeddings based on a text corpus.

Word2vec is a two-layer neural network that processes text by "vectorizing" words. Its inputis a text corpus, and its output is a set of vectors. Feature vectors that represent words in that corpus. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. As the name implies, word2vec represents each distinct word with a particular list of numbers called a vector. The vectors are chosen carefully such that a simple mathematical function (the cosine similarity between the vectors) indicates the level of semantic similarity between the words represented by those vectors.

The General Flow of the Algorithm

- Step-1: Initially, we will assign a vector of random numbers to each word in the corpus.
- Step-2: Then, we will iterate through each word of the document and grab the vectors of the nearest n-words on either side of our target word, concatenate all these vectors, and then forward propagate these concatenated vectors through a linear layer + SoftMax function, and try to predict what our target word was.
- Step-3: In this step, we will compute the error between our estimate and the actual target word and then backpropagate the error, and then modifies not only the weights of the linear layer but also the vectors or embeddings of our neighbor's words.
- Step-4: Finally, we will extract the weights from the hidden layer and by using these weights encode the meaning of words in the vocabulary.

Word2Vec model is not a single algorithm but is composed of the following two pre-processing modules or techniques:

Continuous Bag of Words (CBOW) model:

The aim of the CBOW model is to predict a target word in its neighborhood, using all words. Topredict the target word, this model uses the sum of the background vectors. For this, we use the pre-defined window size surrounding the target word to define the neighbouring terms that are taken into account.

Skip Gram:

The continuous skip-gram model learns by predicting the surrounding words given a current word. The Skip-Gram model is trained on n-gram pairs of (target word, context word) with a token as 1 and 0. The token specifies whether the context words are from the same window or generated randomly. The pair with token 0 is neglected. the skip-gram model is the exact opposite of the CBOW model**.**

- **LSTM**

Long Short-Term Memory Networks, most commonly referred to as "LSTMs," are a unique class of RNN that can recognize long-term dependencies. They are currently frequently used and perform incredibly well when applied to a wide range of issues. Intentionally, LSTMs are created to prevent the long-term reliance issue. They don't struggle to learn; rather, remembering information for extended periods of time is basically their default behavior. Although the repeating module of LSTMs also has a chain-like topology, it is structured differently. There are four neural network layers instead of just one, and they interact in a very unique way.

Input gates control writing operations, input modulating gates decide how much to write in, forget gates carry out erase/remember operations, and output gates decide what output to produce from the cell memory. The input gate regulates the input stream of data to the memory cell, and the output gate controls the output stream of data from the memory cell to other LSTM blocks, to describe how the gates function.

4 Results

```
score = model.evaluate(x_test, yval, batch_size=32)
print()
print("ACCURACY:",score[1])
print("LOSS:",score[0])
```

```
35/35 [==============================] - 4s 101ms/step - loss: 0.0727 - accuracy: 0.9812

ACCURACY: 0.9811659455299377
LOSS: 0.07265845686197281
```

Fig 2. Result of various models with the proposed model

The neural network deep learning algorithms that we used is LSTM(long-short term memory). This algorithms worked well on trip advising. We got 98% accuracy folds. LSTM has four layers 1)LSTM layer-1 2)LSTM layer-2 3)drop out layer 4)dense layer.

5 Conclusion

The previous or existing trip advising systems used traditional text-based machine learning models. The results highly rely on the crafted extracted features. The performances areunstable when advising trips.

So, we propose a deep learning model based on the LSTM(Long short-term memory) method of trip advising systems. The neural network deep learning algorithm that we usedis LSTM(long-short-term memory). This algorithm worked well on trip advising systems. We got 98% accuracy.

6 References

1. A. A. Wadhe and S. S. Suratkar, "Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques," *2020 International Conference on Industry 4.0 Technology (I4Tech)*, 2020, pp. 1-6, doi: 10.1109/I4Tech48345.2020.9102673.

2. https://www.frontiersin.org/articles/10.3389/fcomp.2021.775368/full

3. Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216-225.

4.https://www.researchgate.net/publication/362886308_Word2Vec_on_Sentiment_Analysis_with_Synthetic_Minority_Oversampling_Technique_and_Boosting_Algorithm/citation/download

5. https://www.researchgate.net/publication/312421107_Travel_time_prediction_with_LSTM_neural_network

6.https://www.researchgate.net/publication/332825465_Sentiment_Analysis_Approach_Bas ed_on_N-gram_and_KNN_Classifier

7. https://www.actonscholars.org/sentiment-analysis-with-knn-algorithm/

8. AMIRA, Siti Azza; IRAWAN, M. Isa. Opinion Analysis of Traveler Based on Tourism Site Review Using Sentiment Analysis. **IPTEK The Journal for Technology and Science**, [S.l.], v. 31, n. 2,p. 223-235, may. 2020. ISSN 2088-2033. Available at:

9. TY - JOUR AU - Nawangsari, Rizka AU - Kusumaningrum, Retno AU - Wibowo, Adi PY - 2019/01/01 SP - 360 EP - 366 T1 - Word2Vec for Indonesian Sentiment Analysis towards Hotel Reviews: An Evaluation Study VL - 157 DO - 10.1016/j.procs.2019.08.178 JO - Procedia Computer Science ER

10. https://iopscience.iop.org/article/10.1088/1742-6596/1175/1/012248