A Deep Learning Approach for Sentiment Prediction Using LSTM Networks

Dr. Tejashree T. Moharekar
Assistant Professor
Yashwantrao Chavan School of Rural Development, Shivaji University, Kolhapur
ttm.50649@unishivaji.ac.in

Dr. Urmila R. Pol
Associate Professor
Department of Computer Science, Shivaji University, Kolhapur
urp_csd@unishivaji.ac.in

Dr. Tejpal J. Moharekar
Assistant Professor
Department of Commerce and Management, Shivaji University, Kolhapur.
tjm.50606@unishivaji.ac.in

ABSTRACT

The objective of sentiment analysis is to discover the positive and negative reviews. Hotel reviews have been analyzed in this paper, we have developed a binary text classifier. To prepare the data for the text classifier, we employ several natural language processing (NLP) preprocessing techniques and rely on LSTM layers. There are more than 38932 data samples in the dataset, split evenly between good and negative comments on the hotel's services and general atmosphere. A situation's underlying perspective or sentiment can be uncovered with the use of sentiment analysis. Understanding the tone or motivation of a message through text, audio, or any other medium. Text analysis and natural language processing (NLP) technologies are used to determine the potential subjectivity of data. In order to classify and work with the data more easily, we need to detect, extract, and measure such details from the text.

Keywords: NLP, sentiment analysis, hotel reviews, emotions, text analysis, EDA

INTRODUCTION

Natural language processing, also known as NLP, is the foundational principle upon which sentiment analysis is constructed. Understanding a variety of concepts from a piece of text is the focus of natural language processing, which is a subfield of sentiment analysis that encompasses the broader field of sentiment analysis. The field of artificial intelligence known as natural language processing (NLP) focuses on texts and aims to give computers the ability to comprehend and draw meaning from written material. It includes everything from counting the number of words to a machine writing a story that is indistinguishable from texts written by humans and can be used for tasks such as virtual assistant, query solving, creating and maintaining human-like conversations, summarizing texts, spam detection, sentiment analysis, and other similar tasks. The field of sentiment analysis can be broken down into a wide variety of subfields according to a wide range of criteria. It is possible to categorize it as document-level sentiment analysis, sentence-level sentiment analysis, and sub-sentence level or phrase-level sentiment analysis, depending on the scope. Models of LSTM, also known as extended short term memory, are utilized in order to resolve these issues. LSTMs are a subcategory of rnn models that are distinguished by their capacity to acquire knowledge of long-term dependencies. They are trained to remember information over the long term. The inclusion of a memory cell, which is an additional recurrent state, and several gates, one on each cell, which are responsible for controlling the flow of information into and out of the memory cell, is what distinguishes LSTM models from other types.

LITERATURE REVIEW

Luxury hotels with pricey facilities offer more premium service and achieve greater customer loyalty than mid-range hotels and budget hotels. Luxury hotels are distinguished by their expensive facilities (Liu et al., 2017). When high-end furnishings and facilities become the norm and usual in practically all luxury hotels, guests tend to become more particular about the finer points of their stay and demand more attentive care. When compared to low- and middle-priced hotels, luxury hotels place a greater emphasis on the feedback and opinions of its clients. This is due to the fact that clients of luxury hotels are significantly more loyal than clients of other types of hotels (Knutson et al., 1993) and a higher level of brand loyalty can lead to an increase in sales as well as a greater intention to buy (Chaudhuri & Holbrook, 2001). Hotels should pay attention to even the smallest requests made by guests and be prepared to fulfill them in advance if they want to keep their consumers coming back. In this scenario, managers have an obligation to place a high priority on receiving input from customers. There are a variety of methods that luxury hotels use to gather feedback from guests. The conventional approach is gathering feedback in the form of written comments on a customer book, examining the results of customer questionnaires, and having direct conversations with customers.

Customers today have the option of providing comments via the hotel's app on their cell phones, as well as giving reviews on the hotel's official website and hotel review website. They are able to submit feedback regarding their stay at the hotel and rate various features of the establishment, such as the hospitality of the employees, the convenience of the location, and the quality of the breakfast. Apps and websites provide hotels with opportunities to engage in two- way communication with guests and to amass the feedback essential to enhancing the standard of their services. Customers have access to a forum where they may voice their opinions and suggestions regarding the hotels they have been at on websites that specialize in tourism and hotel reservations, such as Agoda, TripAdvisor, and Booking.com. Customer feedback on hotels provides insight into the products and services that customers value (Xie, Zhang, & Zhang, 2014; Giachanou & Crestani, 2016).

For instance, the images of hotel rooms that reviewers submit present the sanitary conditions of the hotel, and the remarks concerning parking demonstrate whether or not this hotel is accommodating to travelers who drive themselves. As a result, traveler reviews are a crucial source of tourism information, and they also play a significant part in the selection process for hotels. In addition, several websites that assess tourism destinations also analyze user reviews and then rank or recommend hotels for travelers depending on the results of the analysis. When it comes to collecting, comprehending, and responding to the concerns of hotel guests, online review is also a very important tool for hotel management. Because of developments in many forms of technology as well as shifts in the preferences of consumers (Nieves & Segarra-Ciprés, 2015), The management of hotels need keen observation skills in order to keep up with the ever-evolving expectations of their guests. In this scenario, management can benefit from using an online review platform that includes real-time feedback from customers (Mauri & Minazzi, 2013). Customer evaluations are becoming an increasingly prominent subject of research in the field of business as a consequence of the value that they provide to hotels as well as to travellers. In the recent years, a significant number of researchers have been conducting sentiment analysis on customer evaluations (Geetha, Singha, & Sinha, 2017). When conducting this kind of research, various approaches are typically taken. Conducting surveys is one of the most used methods. In order to determine the factors that contribute to customers' trust in online reviews, researchers create online questionnaires for customers to fill out (Filieri, Alguezaui, & McLeay, 2015).

Researchers offer guidance to hotels based on the results of the study regarding how these reviews can be used by hotels to increase both their income and the quality of their customer service. The econometric method has also been utilised by a number of research to investigate the financial value of customer evaluations (Xie, Zhang, & Zhang, 2014).

This strategy relies on statistical and mathematical models to make predictions about the utility and credibility of internet reviews. Insights into hotel administration can be gained through any one of the aforementioned approaches. Traditional methods, on the other hand, have a difficult time making full use of the huge volumes of data that can be provided by online reviews.

Therefore, technologies such as opinion mining and sentiment analysis appear to deal with scenarios involving large amounts of data (Liu, 2012). A number of research have made use of various computer technologies in order to extract emotions and analyse textual aspects from online customer evaluations (Barreda & Bilgihan, 2013). Crotts et al. (2009) conducted research on three hotels located in New York City and compiled a list of fifteen characteristics associated with each establishment. They ranked all of the characteristics and discovered that customers were most dissatisfied with the pricing and the staff. It has been determined that using a tool that is supported by a computer tends to lessen interpretation bias and more thoroughly identify products and service features. At the moment, review data for a large scale analysis is obtained from the Internet, and researchers use an NLP technique in order to carry out sentiment analysis (Shi & Li, 2011).

SENTIMENT ANALYSIS

Using a specialized Natural Language Processing tool called Sentiment Analysis, often known as opinion mining, we can determine whether or not a set of data carries a positive, negative, or neutral tone. There are a few ways to go about this, the most common being machine learning and lexicon-based strategies. Sentiment analysis has several applications, including but not limited to enhancing the customer service process, lowering employee turnover, and developing superior products. Sentiment analysis is used by virtually every large and expanding company today to streamline operations and increase revenue. From facial expression analysis to text categorization, sentiment analysis has many uses.

One of the most common applications of sentiment analysis is in the field of business intelligence, where it is used to assist hotels track how their guests feel about the services they provide. It's also used by several social networking sites to determine whether or not a post's sentiment is particularly strong or aggressive, at which point the message is either removed or hidden.

Business executives have found it more useful to use feedback automation tools since the rise of ecommerce and social media. The retail industry is one where sentiment analysis is commonly used. Customers are increasingly vocal about their wants, needs, preferences, and complaints in the digital realm, which is driving the rise in popularity of sentiment analysis in the retail and technology sectors.

Sentiment analysis is a useful tool for entrepreneurs looking to gauge consumer reaction to their products. By monitoring customer reactions to advertisements, shop layouts, and in-store merchandise in real time, retailers and tech businesses may better cater to their customers' needs.

Using a sentiment analysis tool, we could immediately assess how satisfied consumers were with the product as a whole. Insurance businesses can use sentiment analysis to identify recurring themes in customer feedback, claim details, settlement letters, and other documents.

OBJECTIVE OF THE STUDY

The end goal of the research is to construct a hotel review prediction model that is capable of reliably forecasting hotel reviews, regardless of whether those reviews are good or negative.

EDA places a stronger emphasis on addressing missing data and changing variables as appropriate, in addition to analyzing the assumptions that are important for the process of model fitting. In the EDA report, the number of rows and columns in the data set, as well as any missing data, the types of data, and a preview of the feature's significance, are briefly described.

DATA AND METHODOLOGY

The dataset on Hotel Reviews that can be found on Kaggle provides us with the information that we make use of. There are a total of 38932 unique data points for numerous hotels. In addition to the reviews of hotels, the collection includes information regarding the categorization of the reviews as either positive or negative.

We will be utilizing EDA in order to perform analysis on this dataset. The dataset, in addition to the values that correlate to it, is shown in the following table.

	text	label
0	The room was kind of clean but had a VERY stro...	Negative
1	I stayed at the Crown Plaza April -- - April -...	Negative
2	I booked this hotel through Hotwire at the low...	Negative
3	Stayed here with husband and sons on the way t...	Positive
4	My girlfriends and I stayed here to celebrate ...	Negative
...	...	...
38927	We arrived late at night and walked in to a ch...	Positive
38928	The only positive impression is location and p...	Negative
38929	Traveling with friends for shopping and a show...	Negative
38930	The experience was just ok. We paid extra for ...	Negative
38931	The Westin is a wonderfully restored grande da...	Positive

38932 rows × 2 columns

Figure 1: Dataset on Hotel Reviews

DATA PREPROCESSING

This dataset uses categorical labels for its features. Data in numeric form is the only kind that machines can comprehend. Therefore, we need to use the factorise() technique to transform the category variables to their numeric equivalents. This gives back an array of numeric values as well as an index of the categories.

```
sentiment_label = df.label.factorize()
sentiment_label

(array([0, 0, 0, ..., 0, 0, 1], dtype=int64),
 Index(['Negative', 'Positive'], dtype='object'))
```

Figure 2: Factorization

It would be beneficial for us to convert our text data into a format that can be understood by our machine learning model. In essence, we are going to have to transform the text into a list of vector embedding. The connection that exists between the words in the text can be elegantly represented through the use of word embedding. In order to accomplish this, we will first assign a one-of-a-kind number to each of the one-of-a-kind words, and then we will replace that word with the number that we assigned. Tokenizing each and every word in the document with the assistance of the Tokenizer. Tokenization is the process of reducing all of the words and phrases in a given text to smaller units known as tokens.

```
from tensorflow.keras.preprocessing.text import Tokenizer
tokenizer = Tokenizer(num_words=5000)
tokenizer.fit_on_texts(df_new)
```

Figure 3: Tokenization

The fit on texts() method establishes a connection between the words and the numbered identifiers that are assigned to them. Within the Tokenizer, there is an attribute called word index that stores this association in the form of a dictionary. Using the text to sequence() method, we will now replace the words with the numbers that have been allocated to them.

```
encoded_docs = tokenizer.texts_to_sequences(df_new)
```

The length of the sentences included in the dataset does not vary uniformly from one to the next. To ensure that each sentence is the same length, we should use padding.

```
from tensorflow.keras.preprocessing.sequence import pad_sequences
padded_sequence = pad_sequences(encoded_docs, maxlen=200)

padded_sequence

array([[   0,    0,    0, ...,  172,  100,  267],
       [   0,    0,    6, ...,   63,   18,  738],
       [  14,    6,   19, ...,  172,   42,  266],
       ...,
       [ 720,   21,  482, ...,   34,   63,   72],
       [   0,    0,    0, ..., 1653,    6,  755],
       [   0,    0,    0, ...,  295,   11,  212]])
```

BUILDING TEXT CLASSIFIER

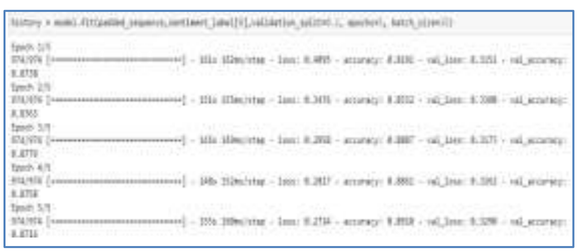
In the machine learning model, we make use of LSTM layers in order to do sentiment analysis projects. An embedding layer, an LSTM layer, and a Dense layer make up the architecture of our model. The Dense layer is located at the very end. We included the Dropout mechanism in-between the LSTM layers to prevent the problem of overfitting the data. Long Short-Term Memory Networks is what the abbreviation LSTM stands for. It is a form of Recurrent Neural Networks that has been modified. The usage of sequential data, such as text and audio, is the most common application for recurrent neural networks.

During the process of generating an embedding matrix, the meaning of each word and the calculations associated with it are often saved. These are referred to as hidden states. When dealing with time-series data, LSTM networks perform quite well. Dropout is one of the methods that are used to regularize the population. It is utilized in order to prevent overfitting. As part of the dropout mechanism, we choose to eliminate certain neurons at random. The layer accepts an argument, which is a number in the range of 0 to 1 that denotes the probability that the neurons will be dropped. This results in a reliable model that is free from overfitting.

```
model = Sequential()
model.add(Embedding(vocab_size, embedding_vector_length, input_length=200))
model.add(SpatialDropout1D(0.25))
model.add(LSTM(50, dropout=0.5, recurrent_dropout=0.5))
model.add(Dropout(0.2))
model.add(Dense(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy',optimizer='adam', metrics=['accuracy'])
print(model.summary())

Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_1 (Embedding)     (None, 200, 32)           824576

 spatial_dropout1d_1 (Spatia  (None, 200, 32)          0
 lDropout1D)

 lstm_1 (LSTM)               (None, 50)                16600

 dropout_1 (Dropout)         (None, 50)                0

 dense_1 (Dense)             (None, 1)                 51

=================================================================
Total params: 841,227
Trainable params: 841,227
Non-trainable params: 0
_________________________________________________________________
None
```

TRAINING SENTIMENT ANALYSIS MODEL

Train the sentiment analysis model for five iterations using the entire dataset, with a batch size of 32 records and a validation split of 20 percent. What the output looks like while training is shown below:



On the training set, the sentiment analysis model achieved an accuracy of 89%, while on the test set, it achieved an accuracy of 87%.

VISUALIZATION OF MODEL METRICS

Utilizing the matplotlib package to create plots of the created model's metric values The plot of the accuracy allows us to observe that the model could definitely be trained a little bit more, as the trend for accuracy on both datasets is still rising during the last few epochs, despite the fact that the plot shows that the accuracy has been climbing for quite some time.

We can also observe that the model has not yet become very proficient with the training dataset, as it demonstrates skills that are comparable across both datasets.

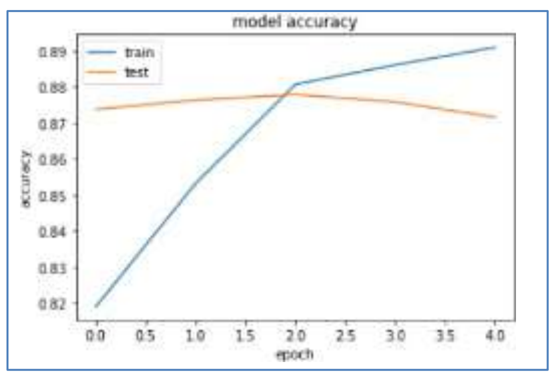


Figure 4: Accuracy Plot

The plot of the loss allows us to observe that the model achieves results that are comparable whether it is applied to the train dataset or the validation dataset (labeled test). If these parallel plots begin to deviate from one another in a regular manner, it may be a hint that training should be stopped at an earlier epoch.

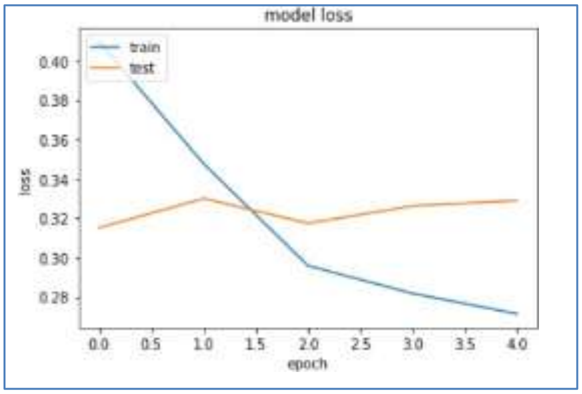


Figure 5: Loss Plot

SENTIMENT CLASSIFICATION PREDICTION MODEL OUTPUT

Define a function that takes a text as input and outputs its prediction label.

```
def predict_sentiment(text):
    r = tokenizer.texts_to_sequences([text])
    r = pad_sequences(r,maxlen=200)
    prediction = int(model.predict(r).round().item())
    print("Predicted label: ", sentiment_label[1][prediction])

Review1 = "Hotel was quite good to stay with family"
predict_sentiment(Review1)

1/1 [==============================] - 0s 117ms/step
Predicted label:  Positive

Review2 = "The service was not proper"

predict_sentiment(Review2)

1/1 [==============================] - 0s 103ms/step
Predicted label:  Negative
```

CONCLUSION

As network technology gets better, more and more people like to talk about hotels on the Internet. And the emotional classification of hotel reviews can not only show how satisfied the public is with the hotel, but it can also help businesses make suggestions to customers about how to improve service and better meet their needs. So, it's important to look at how people feel in hotel reviews. Emotional analysis is figuring out what kind of emotion the text is about based on the comments, whether they are positive or negative. But computers are not like humans. LSTM has been used to make a successful model for analyzing how people feel. In this machine learning model, we made a binary text classifier that divides hotel reviews into positive and negative reviews based on how they make you feel. During validation, we got more than 89% right. The best way to decide if a hotel is right for you is to find out what other people who have stayed there have to say about it. This is where the task of analyzing the tone of hotel reviews can help hotel businesses decide if a hotel is good for a trip or not. This is an interesting piece of information that helps businesses in the Hotel industry figure out how their customers feel about their service.

REFERENCES

Liu, M. T., Wong, I. A., Tseng, T. H., Chang, A. W. Y., & Phau, I. (2017). Applying consumer-based brand equity in luxury hotel branding. Journal of Business Research, 81, 192-202.

Knutson, B., Stevens, P., Patton, M., & Thompson, C. (1993). Consumers' expectations for service quality in economy, mid-price and luxury hotels. Journal of Hospitality & Leisure Marketing, 1(2), 27-43.

Chaudhuri, A., & Holbrook, M. B. (2001). The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty. Journal of marketing, 65(2), 81- 93.

Xie, K. L., Zhang, Z., & Hang, Z. (2014). The business value of online consumer reviews and management response to hotel performance. International Journal of Hospitality Management, 43, 1-12.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. ACM Computing Surveys (CSUR), 49(2), 28.

Nieves, J., & Segarra-Ciprés, M. (2015). Management innovation in the hotel industry. Tourism Management, 46, 51-58.

Mauri, A. G., & Minazzi, R. (2013). Web reviews influence on expectations and purchasing intentions of hotel potential customers. International Journal of Hospitality Management, 34(1), 99-107.

Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis. Tourism Management, 61, 43-54.

Filieri, R., Alguezaui, S., & McLeay, F. (2015). Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth. Tourism Management, 51, 174-185.

Barreda, A., & Bilgihan, A. (2013). An analysis of user-generated content for hotel experiences. Journal of Hospitality and Tourism Technology, 4(3), 263-280.

Crotts, J., Mason, P., & Davis, B. (2009). Measuring guest satisfaction and competitive position: Application of stance analysis to blog narratives. Journal of Travel Research, 48(2), 139-151.

Shi, H. X., & Li, X. J. (2011). A sentiment analysis model for hotel reviews based on supervised learning. In 2011 International Conference on Machine Learning and Cybernetics (Vol. 3, pp. 950-954). IEEE