Sentiment Analysis & Neural Networks

An insight into hotel reviews and predictions

Nawaz A.Rahman

Thesis Report 30p

Simon Celinder

Data Scientist

Göteborg | 2 May 2023

# Table of Contents

# List of Abbreviations

NLTK - Natural Language Toolkit

EDA - Exploratory Data Analysis

N\_grams -

LSTM - Long Short-Term Memory

# Background

Sentiment analysis is a technique that has found greater use in a variety of industries, including hotel management. This project's goal is to conduct a sentiment analysis of hotel reviews using NLP tools like LSTM and NLTK. The project will investigate how effectively these tools function in order to categorize hotel reviews as either positively or negatively based on the information provided.

One of the major benefits of using sentiment analysis is that it allows for a quick and efficient analysis of large amounts of data, which may be quite helpful for hoteliers. This project will contribute to a better understanding of how the hotel experience may be improved by examining what factors influence the classification of reviews as positive or negative.

The outcome of the sentiment analysis can also be used to enhance the guest experience at hotels. The most common word combinations and phrases used in negative reviews can be analyzed to help hoteliers pinpoint the most pressing and pressing problem areas. On this basis, hoteliers can enhance their services and, as a result, improve customer satisfaction and community.

In conjunction with this project, LSTM and NLTK will be improved to increase their effectiveness for sentiment analysis of hotel reviews and, as a result, make it possible for hoteliers to quickly and easily identify and fix issues that impair guests' experiences.

# Boundaries

The project's primary goal is to analyze the sentiments of hotel reviews that have been collected from online platforms. Only text data will be used to train and test models with the aid of NLP tools like LSTM and NLTK. The project won't include implementing the results; rather, it will just provide guidance on how to use them to improve the guest experience. The project's timeframe is restricted to the class term, and work will be done on regular workdays.

# Methods

1. Dataset Preprocessing

* Perform exploratory data analysis (EDA) to gain insights about the dataset
* Clean the dataset by removing irrelevant columns, duplicates, and missing values
* Create a new column for the sentiment labels (positive or negative) based on the star ratings and/or text reviews
* Extract the positive column and word count column for further analysis

1. Text Processing

* Use the Natural Language Toolkit (NLTK) library for text processing and cleaning
* Implement a function to tokenize, remove stop words, and perform stemming/lemmatization on the text data
* Use word frequency analysis to identify frequently occurring words and n-grams (bi-grams and tri-grams)

1. Model Preparation

* Split the preprocessed dataset into training and testing sets
* Determine the number of words in the vocabulary and perform visual inspection using boxplots and histograms
* Implement the Interquartile Range (IQR) method to detect and remove outliers
* Tokenize the text data and create a word-to-index dictionary
* Perform padding to ensure that each review has the same length
* Perform one-hot encoding on the sentiment labels
* Reshape the data from 2D to 3D to prepare for LSTM modeling

1. Model Building

* Build an LSTM model using Keras libraries
* Train my own embeddings
* Save the model weights using the ModelCheckpoint function to prevent overfitting
* Train the LSTM model on the training set
* Evaluate the LSTM model's performance using accuracy, precision, recall, and F1-score metrics
* Generate a confusion matrix and summarize the results

1. Model Comparison

* Use the VADER model as a baseline for comparison
* Use the VADER model to make predictions on the test set
* Compare the true positive values against the LSTM and VADER predictions
* Evaluate the LSTM and VADER models based on their scores and prediction accuracy
* Compare the LSTM and VADER models and identify their strengths and weaknesses

Overall, this method involves preprocessing the dataset, cleaning the text data, preparing the data for modeling, building the LSTM model, evaluating the model's performance, and comparing it to the VADER model. This approach can provide a comprehensive analysis of sentiment analysis using LSTM and VADER models.

## Implementation