Sentiment Analysis on Hotel Reviews

Siva Teja Segu , Pavansai Pottimuthi, Ajith Reddy Busipally, Sreehari Revuri
Kent State University

I. Introduction

Sentiment analysis plays a key role in the field of Natural Language Processing, and it is a technique that extract emotions from the raw texts. Most of the E-Commerce sites like Amazon, Flipkart and Google etc., has wide range of applications built on this. Say, for example think about google translator application. It performs brilliantly in understanding, analyzing and translating the data. Also, it is effectively used on social media post and customer reviews in order to know the opinion of the customers whether they are happy or not with product, service and other factors which will play key role in enhancing their businesses.

A. Background

Sentiment analysis, commonly referred to as opinion mining, is a process that uses machine learning, statistical methods, and natural language processing to recognize and extract subjective information from textual data, such as views, attitudes, and emotions. To ascertain the general attitude or sentiment that is being represented in a text is the main objective of sentiment analysis.

A hotel review is a written assessment of a guest's stay at a hotel that is often published on an internet platform such as a travel website or social media. Hotel evaluations may include details about a guest's experience with the rooms, staff, and facilities, as well as their overall happiness with their stay.

B. Problem Statement

The objective of the project to perform sentimental analysis on a hotel review dataset. Given a review by its customers, we need to predict whether the review is good or bad review in other words say it is positive or negative. For each textual review, we want to predict if it corresponds to a good review (the customer is happy) or to a bad one (the customer is not satisfied).

C. Why is it important

As it enables organizations to analyse massive volumes of unstructured data effectively and economically, sentiment analysis is proven to be a useful tool. As a way to divide reviews, it is becoming more and more popular. It is easy to use and can be occasionally simply adjusted. It gives facts and quantifiable data for upcoming decision-making, and when done well, it delivers value to a business. Sentiment research should be used by businesses that want to improve their goods and services, increase sales, and outwit their rivals.

D. Plans to implement

To swiftly determine whether a review is good or negative, sentiment analysis is required. This paper offers a solution by utilizing the Random Forest Classifier, Word2Vec approach to categorize positive and negative opinion reviews, and by comparing models using preprocessing, feature extraction, and feature selection. Due to their communities' stronger inclination toward data science, such as NLP and Deep Learning for Sentiment Analysis, open-source libraries in programming languages like Python and Java are particularly well suited to creating specialized Sentiment Analysis solutions. But, this demanded a lot of time, money up front, and resources.

II. Literature Review

One of the main tools that a machine may use to comprehend human psychology is sentiment analysis. In order to be used in domains where people were previously required to identify mood or emotion, this technology is currently the subject of substantial research. It plays a key role in chat bot assistants, and when paired with speech recognition technology, it may also be used to replace people in contact centers.

For NLP, machine learning methods were introduced. Several of the systems created during this time period employed machine learning methods such as decision trees to build systems of hard if-then rules comparable to existing hand-written rules. In NLP, the hidden Markov models employed part-of-speech tagging, and certain statistical models that make probabilistic judgments were developed. For sentiment categorization, there are five different machine learning classifiers: Nave Bayes, K-Nearest Neighbor, Support Vector Machine, Logistic Regression, and Random Forest.[1]

Nandal in this paper classified amazon product reviews for sentiment analysis using SVM(Support Vector Machine) Tool. The study examined how words can shift in meaning depending on the context in which they're used, and how this impacts the overall evaluation of a product and its specific features.[2].

Humera Shaziya classified movie reviews for sentiment analysis using WEKA Tool. They enhanced the earlier work done in sentiment categorization which analyzes opinions which express either positive or negative sentiment [3].

Ahmad Kamal designed an opinion mining framework that facilitates objectivity or subjectivity analysis, feature extraction and review summarizing.He used supervised machine learning approach for subjectivity and objectivity classification of reviews. The various techniques used by him were Naive Bayes, Decision Tree, Multi layer Perception and Bagging. He also improved mining performance by preventing irrelevant extraction and noise.[4].

Orestes Appel used natural language processing (NLP)essential techniques, a sentiment lexicon enhanced with the assistance of SentiWordNet,and fuzzy sets to estimate the semantic orientation polarity and its intensity for sentences, which provides a foundation for computing with sentiments. The proposed hybrid method is applied to three different data-sets and the results achieved are compared to those obtained using Naive Bayes and Maximum Entropy techniques[5].

III. Proposed Model

The research began with an examination of many research and review articles on sentiment analysis, and each publication's summary was prepared by reading and comprehending the document. Examine popular classification techniques such as Nave Bayes, Random Forest, k-nearest neighbor, Decision Tree Induction, and Support Vector Machine.

The information was obtained from Booking.com. This dataset includes 515,000 customer reviews and ratings for 1493 premium hotels throughout Europe. In the meanwhile, the geographical location of hotels is supplied for additional examination.

A. Data preparation and data exploration

Pandas describe method will give some statistical parameters of the data set like count, mean and standard deviation. To get a quick overview of the data set we use the dataframe.info() function. Python is an excellent language for data analysis, due to the solid ecosystem of data-centric Python tools.

B. Data cleaning is the fundamental step in any NLP techniques

- Data Cleaning and preprocessing steps.
- Using gensim module to convert the text to vectors using DOC2VEC function
- By using SentimentIntensityAnalyzer we find the polarity scores and plot word cloud.

C. Classification

Random forest is a supervised learning method. The "forest" it creates is an ensemble of decision trees, often trained using the "bagging" approach. The bagging method's main notion is that combining learning models improves the final output. Random forest has the significant benefit of being applicable to both classification and regression issues.

D. Performance Evaluations

A receiver operating characteristic (ROC) curve shows how well a model can categorize binary outcomes. An ROC curve is created by displaying a model's false positive rate vs its true positive rate for each feasible cutoff value. The area under the curve (AUC) is frequently measured and used as a statistic to demonstrate how well a model can identify data points.

IV. Key Concepts

A. Overview of Dataset

As we have taken dataset fromm booking.com. The csv file contains 17 fields. The description of each field is as below

- Hotel_Address: Address of hotel.
- Review_Date: Date when reviewer posted the corresponding review.
- Average_Score: Average Score of the hotel, calculated based on the latest comment in the last year.
- Hotel_Name: Name of Hotel
- Reviewer_Nationality: Nationality of Reviewer
- Negative_Review: Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
- ReviewTotalNegativeWordCounts: Total number of words in the negative review.
- Positive_Review: Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
- ReviewTotalPositiveWordCounts: Total number of words in the positive review.
- Reviewer_Score: Score the reviewer has given to the hotel, based on his/her experience
- TotalNumberofReviewsReviewerHasGiven: Number of Reviews the reviewers has given in the past.
- TotalNumberof_Reviews: Total number of valid reviews the hotel has.
- Tags: Tags reviewer gave the hotel.
- dayssincereview: Duration between the review date and scrape date.
- AdditionalNumberof_Scoring: There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there.
- lat: Latitude of the hotel
- lng: longitude of the hotel

B. Data Cleaning

1) Bag Of Words (BOW): Raw text is initially reprocessed and the processing and cleaning techniques.

2) Cleaned or preprocessed : Remove all unnecessary special characters, if there are words of other accent like polish, German, Spanish etc. Remove or replace them or add the right Unicode to make them readable for machine.

3) Normalize all Data: Make the data properly in a single case letter, either upper or lower. Preferred lower using .lower() function.

4) Stemming and lemmatization: In this methods used by search engines and chat bots to analyze the meaning behind a word. Stemming uses the stem of the word, while lemmatization uses the context in which the word is being used

5) Term Frequency and Inverse Dense Frequency(TF-IDF): Inverse document frequency looks at how common (or uncommon) a word is amongst the document.

IDF =Log [(# Number of documents) / (Number of documents containing the word)]

Term frequency works by looking at the frequency of a particular term you are concerned with relative to the document.

TF = (Number of repetitions of word in a document) / (# of words in document).

For combined TF – IDF = TF(t, d) * IDF(t)

So, using TF and IDF machine makes sense of important words in a document and important words throughout all documents.

6) Word2vec: Word2vec is a combination of models used to represent distributed representations of words in a corpus C. Word2Vec (W2V) is an algorithm that accepts text corpus as an input and outputs a vector representation for each word.

V. Infrastructure and Libraries

A. NLTK

The popular open-source library Natural Language Toolkit (nltk) makes working with human language data in Python

straightforward. It provides a wide range of NLP methods, including as sentiment analysis, part-of-speech tagging, tokenization, stemming, lemmatization, and named entity recognition, among others.

B. SentimentIntensityAnalyzer

It is a rule-based sentiment analyzer, and depending on their semantic orientation, sentences are frequently labeled as either positive or negative. Lastly, we use the polarity scores technique to determine the emotion.

C. Gensim

The gensim library is an open-source Python library for topic modeling and similarity detection in large and complex text datasets.The Gensim algorithms Word2Vec, FastText, Latent Semantic Indexing where it automatically recognize the semantic structure of documents by examining statistical occurrences patterns within a corpus of training texts. There is no need for human input because these algorithms are unsupervised.

D. Google Colab

Google Colab is a free to use Jupyter notebook environment where we can create and run Python code in a web browser. By making GPU-accelerated computing resources and pre-installed machine learning libraries accessible, it is made to encourage collaborative work and study.It is an effective tool for projects involving data science and machine learning, especially those that call for access to huge datasets and substantial processing capacity.

E. GPU machine

we are using GPU machines which are available from a variety of cloud computing providers, such as Amazon Web Services, Microsoft Azure, and Google Cloud. By using GPU machine we can greatly accelerate the training and inference of machine learning models, reducing the time required to train and test models.It is expensive for long term project.

VI. References

1) Sarah Anis, Sally Saad and Mostafa Aref [2020]. Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques. Part of the Advances in Intelligent Systems and Computing book series (AISC,volume 1261).
2) Nandal, N., Tanwar, R. and Pruthi, J. Machine learning based aspect level sentiment analysis for Amazon products. Spat. Inf. Res. 28, 601–607 (2020).
3) Shaziya, Humera, G. Kavitha, and Raniah Zaheer. ”Text categorization of movie reviews for sentiment analysis.” International Journal of Innovative Research in Science, Engineering and Technology 4.11 (2015): 11255-11262.
4) Kamal, Ahmad. Review mining for feature based opinion summarization and visualization. arXiv preprint arXiv:1504.03068 (2015).
5) Appel, O., Chiclana, F., Carter, J., and Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. Knowledge-Based Systems, 108, 110–124.