Review – Driven Insights into Corporate Health

Likhit Krishna Nagineni*, Suraj Aravind Bollapragada*
* Department of Computer Science and Engineering, GITAM School of Technology
GITAM Deemed to be University.

Received on: 2023-09-13 **Accepted on: 2023-10-30**

Keyword:
Data Analysis
Web scarping
Sentiment Analysis
Data Visualization

ABSTRACT

In today's digital age, online reviews have emerged as a vital source of information for consumers seeking to make informed decisions about products and services. Simultaneously, companies have come to recognize the substantial impact of customer feedback on their reputation and overall success. This paper presents a study focusing on the innovative approach of assessing a company's health by analyzing its reviews. The study leverages natural language processing (NLP) techniques and sentiment analysis to analyze large volumes of reviews collected from various online platforms. The objective is to extract valuable insights from customer opinions, sentiments, and feedback to gauge the company's overall health, with a particular emphasis on identifying potential areas of improvement.

Corresponding Author: Email: nlikhitkrishna9@gmail.com

INTRODUCTION

With the significant impact of social media on personal and organizational reputation, understanding the sentiment of social media posts, comments, and interactions has become essential for individuals and businesses alike [1] . Sentiment analysis automatically determines the sentiment or emotional tone text or speech conveys. In social media, sentiment analysis can provide valuable insights into public perception, customer feedback, and brand reputation [2]. By analyzing the sentiments expressed in social media content, individuals and organizations can gauge the overall sentiment trends, identify potential issues, and take appropriate actions to maintain or enhance their online presence.

The study introduces a novel framework encompassing data collection, preprocessing, sentiment classification, and topic modelling. By applying this comprehensive methodology to diverse industries and businesses, the research aims to generate quantifiable metrics for measuring a company's reputation, customer satisfaction, and brand loyalty. By understanding the key drivers of positive and negative sentiment, companies can refine their products, services, and customer engagement strategies to align better with consumer expectations. Moreover, this analysis can serve as an early warning system, alerting businesses to emerging issues and allowing them to proactively address potential crises before they escalate.

The results of this study hold significant implications for both companies and consumers. For businesses, the ability to proactively monitor and manage their online reputation can directly impact their bottom line, leading to improved customer retention and acquisition. For consumers, the research provides a valuable tool to make more informed choices, fostering greater transparency and accountability in the marketplace. This paper presents a pioneering approach to assessing a company's health by analyzing its reviews through NLP and

sentiment analysis. This research contributes to a more informed and interconnected commercial landscape in the digital era by bridging the gap between consumer sentiment and business performance.

ALGORITHM

Web Scraping:

- Import necessary libraries: requests for making HTTP requests, BeautifulSoup for HTML parsing, pandas for data manipulation, and numpy for numerical operations.
- Set the base URL for the airline reviews and specify the number of pages to scrape (pages) and the number of reviews per page (page_size).
- Initialize an empty list called reviews to store the scraped reviews.

Loop Through Pages and Scrape Reviews:

- Use a for loop to iterate through the specified number of pages.
- For each page, construct the URL with the page number and retrieve the HTML content using requests.get().
- Parse the HTML content with BeautifulSoup and extract text data from div elements with the class "text_content."
- Append the extracted reviews to the reviews list.

Create a DataFrame:

- Use NumPy to create a 1-dimensional array from the reviews list and convert it into a pandas DataFrame with a column named 'Reviews.'

Sentiment Analysis:

- Use the Hugging Face Transformers library to load a pre-trained BERT-based tokenizer (AutoTokenizer) and a pre-trained model for sequence classification (AutoModelForSequenceClassification).
- Define a function sentiment_analyser that takes a review, tokenizes it, passes it through the model, and returns the predicted sentiment score (1 to 5).
- Apply the sentiment_analyser function to each review in the 'Reviews' column, limiting the input to the first 512 tokens.
- Create a new column 'Sentiment' in the DataFrame to store the predicted sentiment scores.

Data Analysis and Visualization:

- Create five boolean masks (df1, df2, ..., df5) for each sentiment score.
- Count the number of reviews for each sentiment score and store the counts in a list called Sent.
- Create a pie chart using Matplotlib to visualize the distribution of sentiment scores.

RESEARCH METHOD

British Airways is a 4-star global airline with an average rating of 7.3/10. It is the flag carrier airline in the UK. Their published review dataset derives insights into the company's health and visualises these findings through appropriate data analysis techniques[3] The study focuses on getting not a perfect but a brief idea of where the organization stands in the market, how people feel about it, and whether it should work more on its improvement. This entire data analysis procedure could also be reused to get insights about other businesses. One can use better models than this to get even more optimal results.

Figure 1 shows the flow chart of our study and Figure 2 shows the data frame after scraping the data.

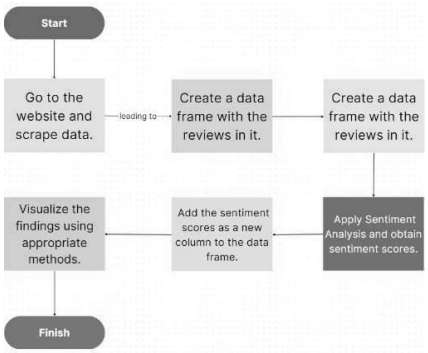

Figure 1: Procedural flow of the study

	Reviews	Sentiment
0	☑ Trip Verified | Having experienced delays a...	1
1	☑ Trip Verified | Travelled to Heathrow to Kal...	2
2	Not Verified | This flight failed at every le...	1
3	Not Verified | Beware of British Airways and ...	1
4	☑ Trip Verified | I flew from Cairo to Heathr...	1
...	...	...
995	☑ Trip Verified | Delivering outstanding cust...	5
996	☑ Trip Verified | This was a night flight New ...	3
997	☑ Trip Verified | Amman to London. Appalling ...	1
998	☑ Trip Verified | Paphos to London Gatwick in...	4
999	☑ Trip Verified | Gatwick to Paphos in Club E...	4

1000 rows × 2 columns

Figure 2: Data frame obtained after scraping and analysing the data

Implementation is done Python using Google Colab with a few libraries and functions imported. The following image illustrates all the import statements that are required.

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
from transformers import AutoTokenizer
from transformers import AutoModelForSequenceClassification
import torch
import matplotlib.pyplot as plt
import matplotlib.ticker as mtick
```

Figure 3: Import statements in Google Colab

Requests and BeautifulSoup are to extract data from the website (Web Scraping). Pandas is for analyzing and manipulating data in a data frame. Numpy enables us to implement mathematical functions on the data. Transformers include all the functions required for NLP purposes. Torch has functions used to build deep neural networks. Matplotlib is essential for data visualization.

[4] used Scrapy for web scraping purposes. BeautifulSoup, on the other hand, is a more lightweight and user-friendly tool than Scrapy. It is ideal for domain-specific tasks like analyzing a business through reviews, even when vast amounts of data are available. Also, Scrapy is a framework created for downloading, editing, and saving data from the web, while BeautifulSoup is a library that helps to pull data from web pages. For the purpose we intend to achieve (analyzing a company's health through reviews), BeautifulSoup is a better fit.

Techniques including Web Scraping, Natural Language Processing (Sentiment Analysis), and Data Visualization are used to achieve the objective. Sentiment analysis is performed using the Bidirectional Encoder Representation (BERT) model[5], which enables the process of each review and assigns a sentiment score.

RESULTS AND ANALYSIS (10 PT)

Using the BERT model from the Transformers library, we could classify the reviews based on the Sentiment Score metric. If the Sentiment Score=1, it implies the least satisfying/terrible reviews. On the other hand, if the score is 5, it implies very positive and satisfied reviews. These are represented in Figure 4 and Figure 5 as a scatter plot and pie charts respectively.

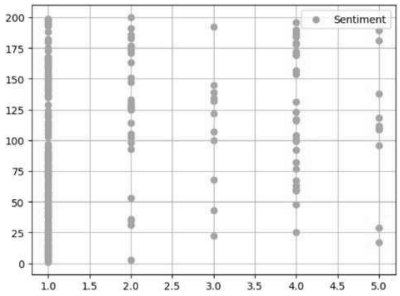

Figure 4: Scatter plot of reviews

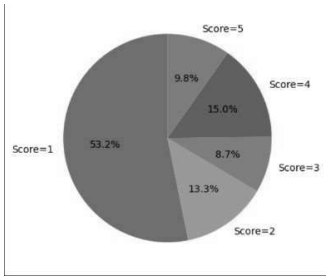

Figure 5: Pie Chart of first 1000 reviews

As evident from the figures 4 and 5 generated after applying the BERT model for the first 1000 reviews, most customers are unsatisfied, i.e., their scores are either 1 or 2. This accounts for a total percentage of 66.5% approximately. 8.7% of the reviews show mediocre satisfaction (Score=3). The remaining 24.8% are happy and satisfied with their experience (Score = 4 or 5)

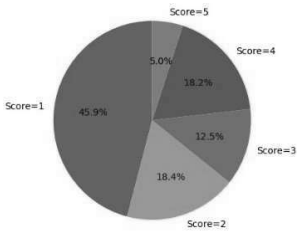

Figure 6: Pie chart of the last 1000 reviews

Figure 6 shows the latest 1000 reviews. The negative reviews have come down to 45.9%. This clearly shows that the organization might have understood the issue for the negative ratings and taken steps to over come it.

The code for the analysis is available in https://github.com/LikhitKrishna2003/Web-Scraping-and-Data-Analysis.git

CONCLUSION

We have successfully extracted data from the review's website, applied Sentiment Analysis to the reviews,

and visualized the inferences. Specialized models like Robustly Optimized BERT (RoBERT) and A Lite BERT (ALBERT) can used, which reduces computational costs.

ACKNOWLEDGEMENTS

We acknowledge British Airways for sharing the review datasets for academic purposes.

REFERENCES

1. A. Veh, M. Göbel, and R. Vogel, "Corporate reputation in management research: a review of the literature and assessment of the concept," Bus. Res., vol. 12, no. 2, pp. 315–353, 2019, doi: 10.1007/s40685-018-0080-4.
2. M. Alzate, M. Arce-Urriza, and J. Cebollada, "Mining the text of online consumer reviews to analyze brand image and brand positioning," J. Retail. Consum. Serv., vol. 67, p. 102989, 2022, doi: https://doi.org/10.1016/j.jretconser.2022.102989.
3. "British Airways webpage." https://www.airlinequality.com/airline-reviews/british-airways (accessed Sep. 06, 2023).
4. D. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019, pp. 450–454, doi: 10.1109/ICECA.2019.8822022.
5. Amiripalli, S. S., Venkatarao, R., Jitendra, M. S. N. V., & Mycherla, N. M. J. (2020). Detecting emotions of student and assessing the performance by using deep learning. International Journal of Advanced Trends in Computer Science and Engineering, 9(2), 1641-1645.
6. Raju, r. k., & Lakshmi, v. a novel implementation of pedestrain detection using hogd and svm algorithms. The Journal of Computational Science and Engineering. pp. 1-7, September 2023.
7. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." 2019.
8. Potharaju, S. P., Sreedevi, M., & Amiripalli, S. S. (2019). An ensemble feature selection framework of sonar targets using symmetrical uncertainty and multi-layer perceptron (su-mlp). In Cognitive Informatics and Soft Computing: Proceeding of CISC 2017 (pp. 247-256). Springer Singapore.
9. Kollu, V. V., Amiripalli, S. S., Jitendra, M. S. N. V., & Kumar, T. R. (2021). A network science-based performance improvement model for the airline industry using NetworkX. International Journal of Sensors Wireless Communications and Control, 11(7), 768-773.
10. Sumathi, A., Kumar, B. S., & Vishnubhatla, S. Advancements in Energy-Efficient Virtual Machine Placement Survey for Cloud Computing. The Journal of Computational Science and Engineering. pp. 9-15, September 2023.