# **Sentiment Analysis of Gassdoor Review**

Bellevue University

Applied Data Science

Madhukar Ayachit

DSC680 - Summer 2022

# Topic

I have decided to work on “Sentiment Analysis of Gassdoor Review”. This project will attempt to decode employee reviews to rate employer or organization.

# Context

Glassdoor is a TripAdvisor for candidates in their job search. Glassdoor offers candidates a chance to look for jobs and read authentic and transparent reviews from employees currently and formerly employed in an organization.

Every organization has their pros and cons which their employees feel that it should be made public so that other people who wants to join this organization make decisions based on reviews from the people.

Glassdoor’s outreach is not limited to potential job seeker, it provides platform for existing employee to express his opinion about his current employer, HR & Employer can monitor their performance by analyzing behavioral patterns recorded in reviews. Similarly, recruiters and Branding team can pay attention to reviews for their strategies.

# Business Problem

# Sentiment score is an important factor all the stake holders: employee, Employer, HR , recruiter and branding team.

This project is an attempt to perform mining on Glassdoor reviews to have a better understanding of what the employee/ex-employee feels about the company. Project sets basic foundation for all the stake holders: employee, Employer, HR , recruiter and branding team to take action to improve performance based on sentiment score of review.

This project presents sentiment analysis of Glasdoor reviews to help to come up with better behavioral strategy.

# Data

Getting Glassdoor data was a challenge. It requires significant communication, and one has to go through stringent process to acquire data. Data is not available in popular sites like Kaggle etc.

I am able to manage a review Dataset from Kaggle for one single employer.

Sample data contains approximately 46K records across 18 columns.

Primary focus being on “ReviewHeadLne”,”Pros” ,”Corns”.

Though dataset contains “rating” , “polarity” but Idea is to derive rating and polarity using “SentimentIntensityAnalyzer”

**About Data**

This is sample Glassdoor review data for “TCS” . It has approximately 46 K rows spread across 18 columns:

* employID
* rating
* reviewDate
* reviewHeadLine
* authorLocation
* autherJobTitle
* pros
* cons
* subRatings
* authorReviewLink
* reviewDateIST
* dayOfWeek
* month
* dayOfMonth
* year
* peher
* polarity
* subjectivity

# Research Questions

Prediction model can help us understand the trend of hidden sentiments in the review. This analysis can help the organization to develop behavioral strategies proactively.

Research question for my work is to predict best sentiment score for all kind of reviews.

**Ethical Implication**

The data available in public domain and carefully chosen to not to violate any data privacy norm.

The way data is formatted and presented, it maintains anonymity of employee and can be used only as sample dataset for academic or research purpose.

**Method**

Solution will be developed in Python using following popular packages/modules:

1. Pandas
2. Seaborn (visualization)
3. Matplotlib (visualization)
4. Sklearn (Logistic Regression)
5. TextBlob
6. WordCloud
7. SentimentIntensityAnalyzer

**Criteria**

“pros” primarily used for positive score

“cons” primarily used to extract negative score

**Exploration & Visualization**

I will be going over basic structure of the data and all the features to get sense of data. I will be probably using simple bar charts to understand all the features and distribution of the data.

**Data Cleanup**

Duplicate and null values are primary target, and I may decide to drop the duplicate and null review phrases.

**Modeling**

I will be following standard practices of data **Standardization** splitting data into **Training/Testing** set as well as looking at **z-score** for **Feature Selection**. I will consider **Logical Regression** , **KNN , SVM and Keras**

I will arrive at final recommendation after evaluating score for each model.

**Challenges/Issues**

1. Based on experience Model iteration may take significant time to test
2. Getting Glasdoor data for difficult and took a while to get a sample data
3. Meaningful utilization of Keras in the context of the model can be tricky

**Analysis**

## **Data Preparation**

Based on the reading, “data scientists usually spend 70% of the project time here, preprocessing and exploring the data”. So, the Data Preparation is important for predictive analytics project. In the Case Study project, we need to make sure the data is suitable and well prepared for model. While preparing the Data, we need to verify the metadata of the given dataset and need to understand very clearly about data patterns, relationships etc.

1. **Acquire and read the data-** Downloading the data directly from the source and reading it.
2. **Clean the data-**Any data from the real world is always messy and noisy. The data needs to be reshaped to aid exploration of the data and modeling to predict the income level.
3. **Explore the independent variables of the data-** A very crucial step before modeling is the exploration of the independent variables. Exploration provides great insights to an analyst on the predicting power of the variable. An analyst looks at the distribution of the variable, how variable it is to predict the income level, what skews it has, etc. In most analytics project, the analyst goes back to either get more data or better
4. Need to describe the dataset **describe**()
5. Need to understand the number of columns and rows shape
6. Need to understand the summary of the data ( **info** )
7. Verify the **Null** values, these missing values impact the model. Hence, we need to treat these missing data.

## **Missing Data Imputation**

As we know, If the variables are treated as continuous or ordinal, leaving these as 0 communicates to the algorithms that 0 is the best answer and if the data were treated as categorical, these could be left as the value 0.

## **Exploratory Data Analysis (EDA)**

We need to perform EDA and that data needs to be cleaned to gain insights that may be useful in improving the performance of our model. Few of the EDA as follows:

* Filling missing information (Null / NA values), replacing values with median
* Encoding of categorical variables
* Dropping of features (or columns)
* Visualization used to show behavior of data and explains insight of data.

## **Feature Selection**

Pros and Cons are the primary source of review. However, there are other features like Location and Job Title to determine the frequency and density of the review.

A part from classification , combination of pros and cons will be used to evaluate sentmiment score.

## **Modeling**

My target variable being binary Logistic Regression was obvious choice.

Importing the necessary libraries needed -

* LogisticRegression for modeling
* Pandas for managing dataset
* Matplotlib.pyplot for visualization
* Seaborn used specially for correlation and confusion matrix visualization

## **Perform prediction & Evaluate Model Performance**

We need to perform the above steps to perfume and evaluate the model performance. The main four main performance metrics used to evaluate the effectiveness of classification models:

* **Accuracy:** test’s ability to correctly predict both classes
* **Precision:** test’s ability to correctly detect positive classes from all predicted positive classes
* **Recall (Sensitivity):** test’s ability to correctly detect positive classes from all actual positive classes.
* **F1 Score:** harmonic mean of precision and recall

## **Challenges**

1. Every iteration was taking significant time to test
2. Had challenge with distribution of the classes
3. Meaningful utilization of Keras in the context of the model

## **Conclusion**

The dependent variable for my dataset is 'Polarity'. The variable contains True and False to determine polarity of the review sentiment.

Based on accuracy output of all the models, I decided to use Logistic Regression regression.

The purpose of the study to extract polarity of the sentiment from review text whether it is positive or negative to convey sentiment of employee.

## **Assumptions**

All the reviews in English language.

## **Limitations**

### Current Implementation is limited to a single data source available in Kaggle. This is mostly cleaner dataset with limited features.

## **Future Uses/Additional Applications**

The data-set we trained here has accuracy is 94%.I conducted hyperparameter tuning Based on the Accuracy indicated in model evaluation I would recommend Linear Regression for this data set.

This project attempts to provide a working model for any text based sentiment analysis especially use cases like "product review" , "twitter review" , "movie review" to know the general perception/opinion of the people/customer/audiences whether they are liking it or rejecting it. Based on the accuracy of the model, the students have a very high chance of predicting which university they can be admitted to for their credentials.

There is an opportunity to improve model performance and determine best model by increasing the size of the data set by collecting more data.

**Recommendations**

**Implementation Plan**

I will be using next two weeks to perform:

1. Data exploration
2. Visual Analysis
3. Data Cleanup
4. Feature selection
5. Target Variable
6. Modeling
7. Testing
8. Conclusion

**Questions from a potential reviewer**

1. Why did you choose this domain?

Sentiment analysis is my field of interest and most likely if opportunity comes, I might do a research work on this topic. During the Master program I did couple of similar assignments too but mostly around social media like twitter or movie review analysis.

In my opinion employees opinion matters for growth of the organization and it should be respected as long as you sense the pulse of the sentiment early stage and act on it to improve overall performance of the organization.

1. What is your data source?

I was trying to get data from Glasdoor but unfortunately it was not an easy process to acquire any companies review data in such a quick time. I have to settle with only available dataset in Kaggle.

1. What is the purpose or business problem you are addressing?

Getting insight of the employee or employees perception on organization. Ideas is to present sentiment analysis to HR/Branding team and management.

1. What are the steps you took to arrive at conclusion?

I followed typical data science cycle from data collection to , exploration, cleanup and arriving at modeling after 80-20 split of sample data into training and testing model. I am relying on Logistic regression due to binary outcome of the target variable.

1. What is your modeling and why?

Outcome is pretty much binary though it mostly falls under typical classification pattern. Though , I have tried to compare score of LR, SVN , KNN , CART before arriving at Logistic Regression.

1. What are the key modules/Libraries being used?

Following key libraries are used in implementation:

1. Pandas to handle dataset
2. Seaborn and matplotlib for graphs and visuals
3. NumPy
4. LogisticRegression for regression
5. TfidfVectorizer for string tokenization before modeling
6. TextBlob for sentiment analysis
7. WordCloud for visualization
8. NTLK
9. Any challenge while importing libraries?

I have dealt with this earlier so it was not challenging this time.

1. What are the key dependent variables and independent variables?

* Combination of “pros” & ‘Corns”
* autherJobTitle
* authorLocation

1. Are use using any visual or graph?

EorldCloud, Bar Chart & Pie chart are being used at various stages of visualizations

1. How do you summarize the result or outcome of your project?

Logistic regression & CART are pretty close to 84% accuracy though I preferred LR to leverage my past experience with it. I did 65-35 split dropping existing rating variables and deriving sentiment score from current execution. Further testing and reptation gave XX score.

**References:**

<https://www.kaggle.com/datasets/ashwini1619/tcsglassdoorreviews>

<https://www.kaggle.com/tanulsingh077/twitter-sentiment-extaction-analysis-eda-and-model/notebook?select=submission.csv>

https://github.com/madhukarayachit/DSC550/blob/main/Ayachit\_Madhukar\_Milestone\_5.ipynb