STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Indeed.com: Review Sentimental Analysis

Course: BIA- 660
Instructor: Dr. Jingyi Sun

Members:
Sahil Mody
Priya Gohil
Viraj Jadhao
Pawanneet Kaur

Contents

- Introduction
- Problem Statement
- Web Scraping
- Dataset Description
- Exploratory Data Analysis
- Sentimental Analysis
- Machine Learning Models

Introduction

- Job searching is a time-consuming procedure for many people all around the world. When applying for a job, one must consider a variety of factors such as the work description, the satisfaction of present employees, the perks, the wage range, and so on.
- The goal of this project is to scrape the prominent Job-Hunting website "Indeed.com" based on the job designation to find mainly the list of companies, user ratings, reviews and descriptions.
- We will categorize the organizations depending on their field and domain, and we will assess existing employee work satisfaction based on the reviews scraped.
- All employee reviews posted on the job board will be examined in order to derive useful insights and employee sentiments. This will assist individuals in picking the companies for which they desire to apply.
- We will do so by various EDA techniques ranging from Histogram, Pie charts, Bar plots and Word Cloud.

Problem Statement

- We've all experienced how difficult and stressful a job search can be. Indeed.com is a popular job-searching website.
- We aim to focus on popular jobs in the market through this project.
- The research part tries to comprehend and analyze the sentiments expressed by workers in job reviews.
- We will examine various work roles and assess the job satisfaction of various firms. This will give people a better idea of what the firm has to offer.

Web Scraping

- For the process of Data collection we used the method of web scraping on the website - indeed.com
- We focused mainly on the library - BeautifulSoup to scrape the reviews of about 150 companies for the two designations, precisely, Data Scientist and Pharmacist
- The attributes scraped are: company name, designation, title, review, date, place and ratings
- This scraped data of company reviews was used further to implement sentiment analysis

Dataset Description

- We visited the website https://www.indeed.com/m/ and looked for the top firms in Data Science, and Pharmacy employment.
- Furthermore, we chose over 150 companies and scraped a list of company names, designation, overall ratings given to the company by employees for a specific sector, the date on which the review was posted, job descriptions of the companies that fall within the relevant areas for the positions, top employee reviews, and the company's location.

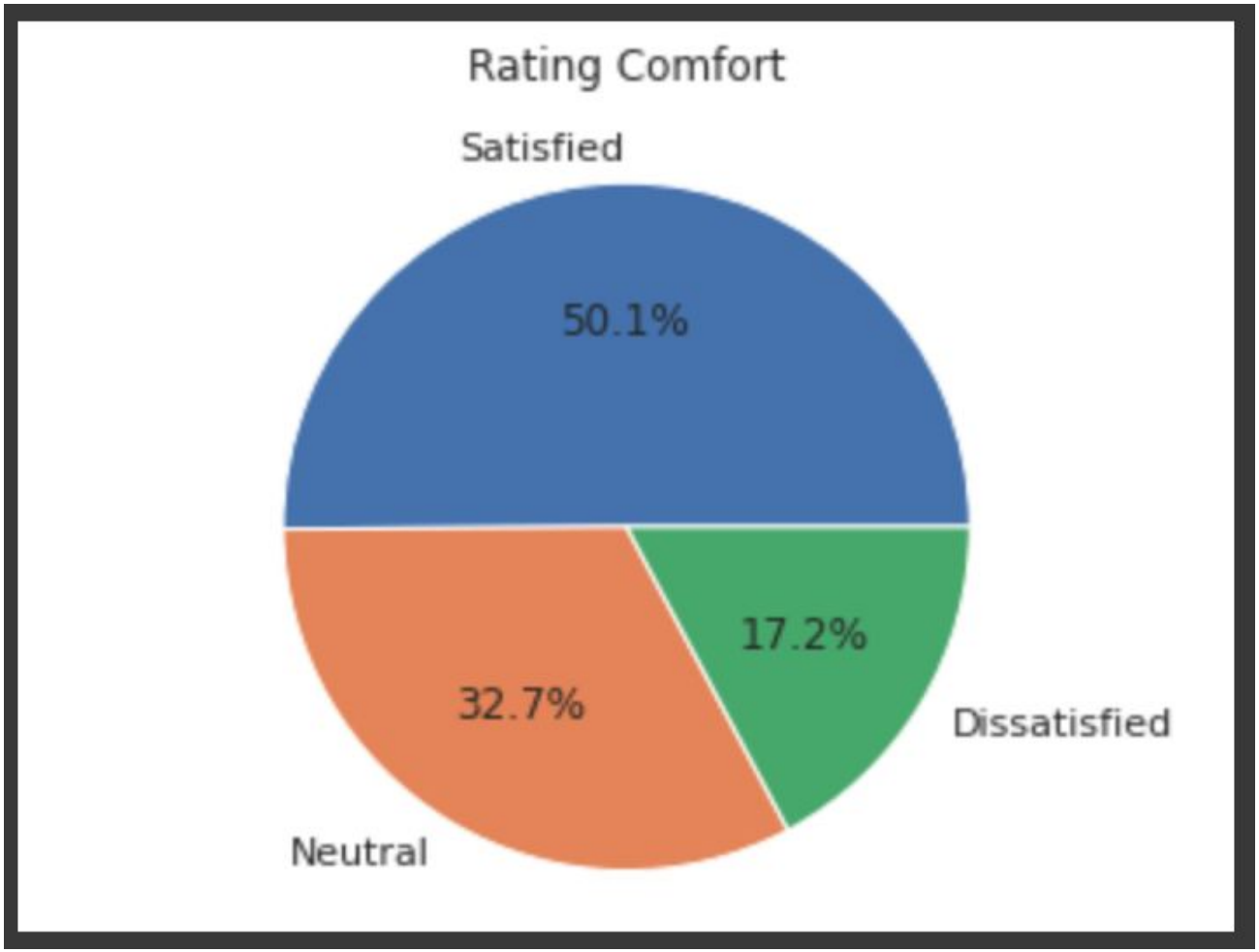
Unnamed: 0	company	Designation	title	review	date	Place	ratings
1	Meta	Software Engineer (Former Employee)	A great place to learn and improve oneself	Working at Facebook exposes you to problems se...	May 27, 2020	New York	5.0
2	Meta	Software Engineer (Former Employee)	Becoming worse	I have joined the company since 2018 and enjoy...	April 24, 2022	California	3.0
3	Meta	PM (Former Employee)	Standard big-tech job, with the pros and cons ...	Great if you're looking for a big tech job, le...	April 24, 2022	Menlo Park, CA	5.0
4	Meta	Program Manager (Former Employee)	Productive and fun environment	It is a great company to work for. I had a pre...	April 21, 2022	Seattle, WA	5.0
5	Meta	Manager (Current Employee)	Love working at Meta!	I've been at Meta (previously Facebook) for a ...	April 7, 2022	Austin, TX	5.0

Exploratory Data Analysis

Word Cloud - Reviews

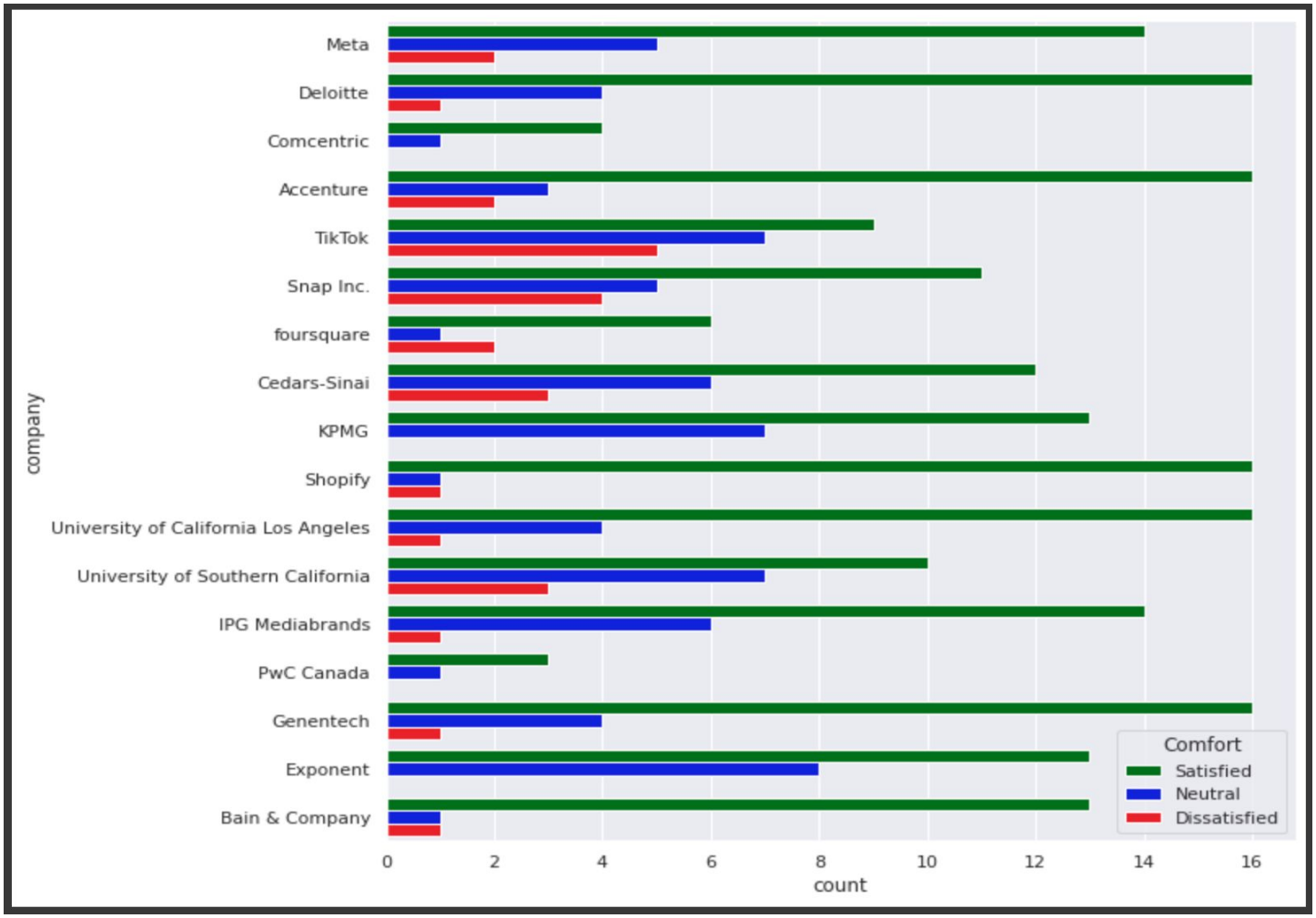
Exploratory Data Analysis

Percentage of Rating Comfort



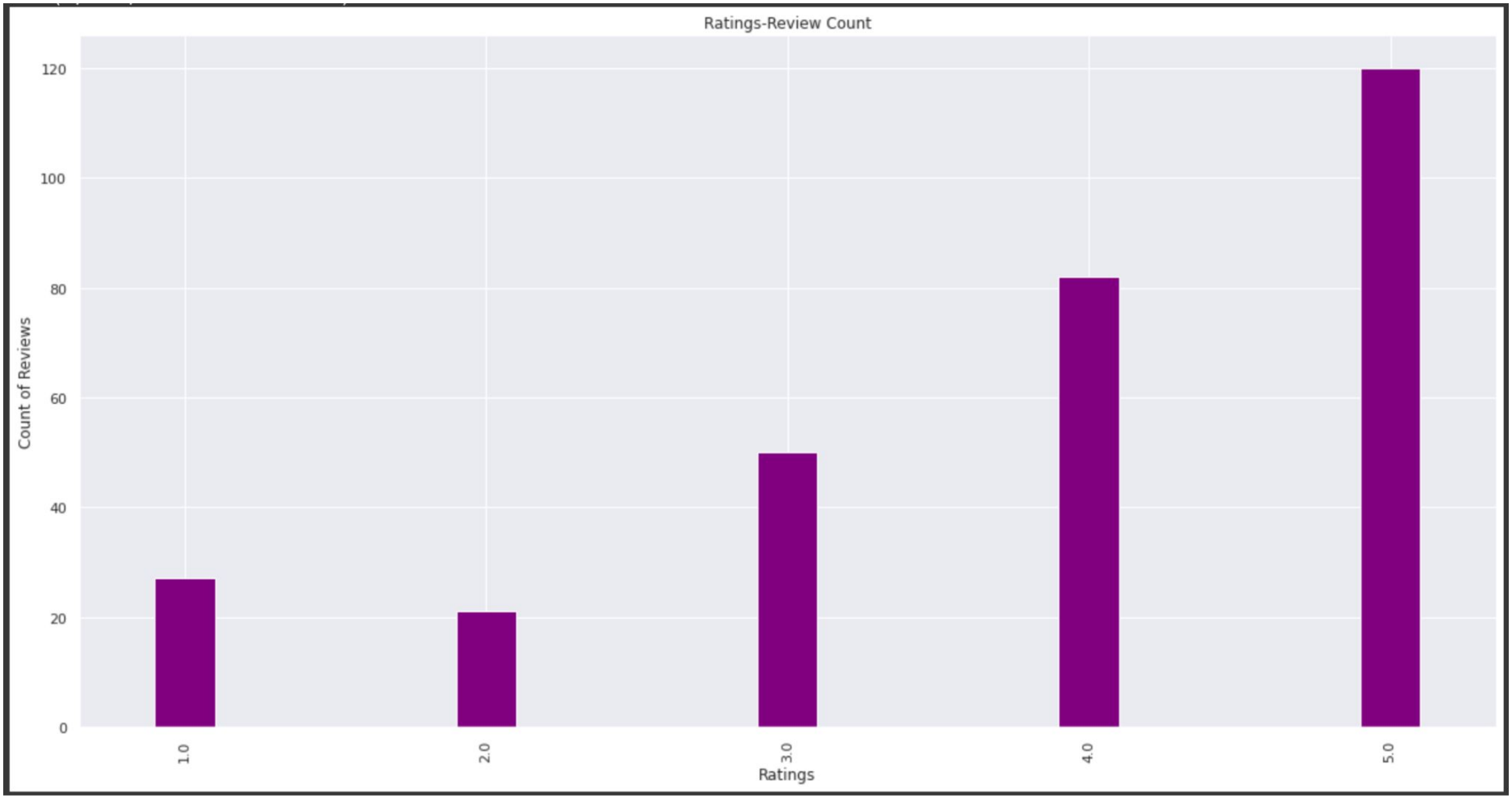
Exploratory Data Analysis

Comfort with respect to Companies



Exploratory Data Analysis

Ratings - Count of Reviews



Sentimental Analysis

- For the purpose of sentimental analysis, we first started with cleaning out all the special characters and punctuation marks using regex.
- We further went ahead to remove all the “stopwords” from the text so that it becomes easier for the system to analyze the text and make the required predictions.
- Post cleaning our textual reviews, using the TextBlob library provided by Python, we classified the reviews into positive, negative and neutral.
  - TextBlob helps the user by returning a polarity value.
  - For our dataset, if the polarity value was less than the sentiment was classified as negative; if the polarity value was equal to zero, the sentiment was neutral and if the polarity was a positive value, the sentiment was assumed to be positive.

Machine Learning

Logistic Regression

- Logistic regression is a supervised Machine Learning algorithm commonly used for classification purposes.
- We used this to predict whether our review was positive or negative.
- Firstly, we used sklearn's Pipeline model, which comprised of **Hashing Vectorizer** and **Logistic Regression**.
  - **Hashing Vectorizer** converts a collection of text documents to a matrix of token occurrences.
    - The main difference between HashingVectorizer and CountVectorizer is that the former does not store the resulting vocabulary or tokens. This makes it extremely suitable for large dataset, such as ours.
- The Logistic model, using the 'saga' solver and 1000 iterations gave us an accuracy of 85%.

Machine Learning

Multinomial Naive Bayes

- Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP).
- It is popularly used for Text data classification, in this case, to classify whether a review was positive or negative.
- This algorithm is based on the popular Bayes Theorem **P(A|B) = P(A) * P(B|A)/P(B).**
- One of the major advantages of this algorithm is that it is highly scalable and appropriate of large datasets.
- In this case, we decided to use Count Vectorizer and we noticed a significant change in the complexity of the program.
- With the help of sklearn library, MultinomialNB, which alpha as 1, also gave us an accuracy of 85.18%.

Machine Learning

Decision Tree Classifier

- Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.
- The intuition behind **Decision Trees** is that you use the dataset features to create *yes/no* questions and continually split the dataset until you isolate all data points belonging to each class.
- With this process you're organizing the data in a **tree structure**.
- In our program we have used a maximum depth of 120 and min_samples_split as 3 to receive an accuracy of 87%.
    - The minimum number of samples required to split an internal node:
    - The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

References

- https://scikit-learn.org/stable/
- https://www.geeksforgeeks.org/python-programming-language/
- Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. “Sarcasm as contrast between a positive sentiment and negative situation”, In EMNLP 2013, pp. 704–714.
- Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. “Sarcasm as contrast between a positive sentiment and negative situation”, In EMNLP 2013, pp. 704–714.
- B. Agarwal, N. Mittal, “Prominent Feature Extraction for Review Analysis: An Empirical Study”, In Journal of Experimental and theoretical Artificial Intelligence, 2014, DOI: 10.1080/0952813X.2014.977830.
- M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, “Lexicon-based methods for sentiment analysis”, Computational Linguistics, v.37 n.2, p.267-307, 2011

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

stevens.edu

Thank You