

A

PROJECT

ON

**A Comprehensive Study of Capgemini Employee Reviews**

Submitted to

MIT ADT UNIVERSITY

SCHOOL OF ENGINEERING & SCIENCE

DEPARTMENT OF APPLIED SCIENCE & HUMANITIES

In partial fulfilment of the requirements for the award of the degree of

MASTER OF SCIENCE

IN

APPLIED STATISTICS (DATA SCIENCE)

Submitted by

Mr. Suraj Ramchandra Jagtap MITU22MSDS0012

Mr. Rohit Mahendra Dixit MITU22MSDS0028

Under the Guidance of

|  |  |
| --- | --- |
| Prof. Rohit Raskar  Guide | Co-guide |

June 2024



SCHOOL OF ENGINEERING & SCIENCE

DEPARTMENT OF APPLIED SCIENCE & HUMANITIES

RAJBAUG, LONI KALBHOR, PUNE-412201

CERTIFICATE

This is to certify that the Capstone Major Project-22MSDS422 entitled

**A Comprehensive Study of Capgemini Employee Reviews**

Submitted by

Mr. Suraj Ramchandra Jagtap MITU22MSDS0012

Mr. Rohit Mahendra Dixit MITU22MSDS0028

is a bonafide work carried out by them, under the supervision of guide and co-guide, it is submitted towards the partial fulfilment of the requirement of MIT Art, Design and Technology University, Pune for the award of the Master of Science in Applied Statistics (Data Science).

|  |  |
| --- | --- |
| Prof. Rohit Raskar  Guide | Co-guide |

|  |
| --- |
| Dr. Pratibha Jadhav  Program Co-ordinator |

|  |  |
| --- | --- |
| Prof. Dr. Haribhau Bhapkar  HOD,  Applied Sciences & Humanities | Prof. Dr. Virendra Shete,  Director,  MIT School of Engineering & Science |

External Examiner

Place: Pune

Date:

**DECLARATION**

We hereby declare that the Project entitled “A Comprehensive Study of Capgemini Employee Reviews” submitted towards the partial fulfilment of the requirement of MIT-ADT University, Pune for the award the Master of Science in Applied Statistics (Data Science) of the is a record of bonafide work carried out by us under the supervision of Guide Name and Co-guide name, Department of Applied Science & Humanities, MIT School of Engineering & Science, Pune.

We further declare that the work reported in this report has been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Pune Student Names and Sign

………………..

(Suraj R. Jagtap)

………………..

(Rohit M. Dixit)

**ACKNOWLEDGEMENT**

I would like to thank all teaching and non-teaching staff and all my colleagues. I would like to express my profound gratitude towards guide Prof. Rohit Raskar and co-guide Prof. Satish Jadhav for his valuable guidance for completion of this project. I am also thankful to Dr. Pratibha Jadhav, Rohit Raskar Sir, for their timely suggestions and encouragements.

I would like to thank Prof. Dr. Haribhau Bhapkar, Head of Department Applied Science & Humanities, MIT Art, Design and Technology University, Pune (India), for providing me the necessary facilities.

**ABSTRACT**

The Capgemini employee review dataset is a treasure trove of insights into employee sentiments and experiences within the organization. This dataset comprises thousands of employee reviews, each offering valuable information on various aspects such as work environment, company culture, management style, career growth opportunities, and overall satisfaction.

Analyzing this dataset can unveil crucial patterns and trends in employee feedback, aiding in understanding the factors influencing employee morale and engagement within Capgemini. With a comprehensive study, several key components can be outlined:

* **Data Collection and Preprocessing:** We collected and cleaned a vast dataset of Capgemini employees’ reviews. The data included Department, Overall ratings and more.
* **Natural Language Processing (NLP):** we can develop an NLP model to extract the keywords from employees’ reviews which can helps us to see the important word based on Tf-Idf score. the main of NLP tool is to deal with text data and find out the meaningful insights.
* **Recommendation System:** Building a recommendation system based on department and place. It uses employee reviews to identify patterns and trends, guiding targeted interventions to address dissatisfaction and reinforce positive work environment. The system may involve policy changes or management practices tailored to each location and each department.
* **Sentiment Analysis:** We performed sentiment analysis on ratings to understand employee’s satisfaction.

**Keywords:**

Data Scientist, Data Analyst, Data Engineer, python, SQL, Exploratory Data Analysis, Recommendation System, Natural Language processing, sentiment analysis, LSTM, Navie Bayes algorithm.

**CONTENTS**

Title Page………………………………………………… I

Certificate ……………………………………………….. II

Declaration ……………………………………………… III

Acknowledgement ………………………………………. IV

Abstract ………….………………………………………. V

Chapter-1 Introduction .……………………………………………. 8-9

Chapter-2 Literature Reviews ..……………………………………. 10

Chapter-3 Objective- .……………………………………………… 11

Chapter-4 Methodology- …………………………………………… 12-36

4.1 Data Description ..…………………………………. 14-15

4.2 Data Preprocessing .………………………………. 16-17

4.3 Data Visualization ….…………………………….. 18-23

4.4 Analysis Using Software ………………………….. 24

Chapter-5 Conclusion and Discussion …………………………….. 37

Chapter-6 Limitation and Future Scope …………………………… 38-39

Chapter-7 Reference ………………………………………………. 40

Annexures: Plagiarism Report …………………………. 41

**Chapter-1**

**Introduction**

The Capgemini Employee Reviews dataset is a comprehensive repository of employee feedback and sentiments within the Capgemini organization. This dataset encapsulates a wide array of perspectives, spanning from January 2018 to March 2022, offering a detailed and nuanced understanding of the employee experience within the company. This dataset serves as a rich source of information on factors influencing employee satisfaction, engagement, and overall workplace dynamics.

At its core, the dataset comprises employee reviews, offering candid assessments of different facets of the Capgemini work environment. From job satisfaction and work-life balance to compensation, career growth opportunities, and company culture, the dataset delves into the multifaceted dimensions that shape the employee experience. Just as weather data provides essential metrics like temperature, rainfall, humidity, and wind speed, this dataset offers key indicators crucial for evaluating organizational health and employee well-being

Each review within the dataset represents a unique perspective, reflecting the diverse experiences and opinions of Capgemini employees across various roles, departments, and locations. By aggregating and analyzing this wealth of feedback, stakeholders gain valuable insights into prevailing trends, patterns, and areas of strength and improvement within the organization. Similar to how urban planners utilize weather data to design resilient infrastructure, HR professionals and organizational leaders can leverage this dataset to inform strategic decision-making and enhance the overall employee experience.

The dataset facilitates a granular examination of employee sentiments over time, enabling longitudinal analysis and trend identification. Researchers and analysts can explore how employee perceptions evolve across different periods, uncovering potential correlations with organizational changes, events, or initiatives. Additionally, the dataset allows for comparisons across different demographic groups, such as tenure, job level, or department, providing valuable insights into potential disparities or areas requiring targeted interventions.

Moreover, the dataset offers opportunities for sentiment analysis and natural language processing (NLP) techniques, enabling the extraction of actionable insights from unstructured text data. Advanced analytics methodologies can uncover underlying themes, sentiments, and sentiment drivers within employee reviews, providing deeper contextual understanding and actionable recommendations for organizational improvement.

Beyond internal stakeholders, the Capgemini Employee Reviews dataset may also be of interest to external parties, such as prospective employees, investors, or industry analysts. Transparent access to employee feedback fosters accountability and trust, showcasing Capgemini's commitment to fostering a positive work environment and continuously improving employee satisfaction and engagement.

**Chapter-2**

**Literature Reviews**

Logistic regression, developed by David Cox in 1958, is a statistical method for binary classification tasks. Its extension, multinomial logistic regression, was introduced in 1972, allowing for multi-class classification. It's widely used in healthcare, economics, and social sciences for flexible and interpretable modeling of categorical outcomes.

1. Ramadhan, Novianty, and Setianingsih's (2017) paper discusses sentiment analysis methodologies, focusing on machine learning techniques like Naive Bayes, Support Vector Machines, and decision trees. They highlight the limitations of Naive Bayes and highlight the potential of Multinomial Logistic Regression for multiclass classification. They emphasize the importance of feature selection and data preprocessing.

The Naive Bayes algorithm, developed by Thomas Bayes in the 18th century, simplifies probabilities of events based on prior knowledge. Its application expanded to pattern recognition and machine learning, particularly in text classification tasks like spam filtering and sentiment analysis due to its simplicity and efficiency.

1. Hasanli and Rustamov's 2019 study examines sentiment analysis techniques for Azerbaijani tweets, focusing on logistic regression, Naive Bayes, and SVM. The review explores previous research on sentiment analysis methodologies, particularly in social media data and Azerbaijani language, and sets the stage for future research.

The Long Short-Term Memory (LSTM) algorithm, introduced by Hochreiter and Schmidhuber in 1997, addresses the vanishing gradient problem in traditional Recurrent Neural Networks. It uses memory cells and gating mechanisms to selectively retain or forget information over long sequences, revolutionizing sequential data processing in fields like natural language processing, speech recognition, and time series prediction.

1. Minaee, Azimi, and Abdolrashidi (2019) discuss sentiment analysis techniques, highlighting the importance of deep learning. They highlight traditional methods like Naive Bayes, Support Vector Machines, and decision trees, but highlight the superior performance of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks in capturing semantic nuances and contextual information. Combining these models improves accuracy and robustness in sentiment classification tasks.

Latent Dirichlet Allocation (LDA), introduced in 2003 by Blei, Ng, and Jordan, is a probabilistic model that identifies latent topics in large text datasets. It assumes documents are mixtures of topics, with topics being distributions over words. LDA builds on previous models, providing a generative approach with a clear probabilistic foundation.

1. Jelodar et al. (2019) discuss the use of natural language processing (NLP) in recommendation systems, specifically the Latent Dirichlet Allocation (LDA) topic model. They highlight its effectiveness in extracting thematic content from textual data, improving relevance and personalization, and integrating it with collaborative and content-based filtering.

**Chapter-3**

**Objectives**

The specific objectives are:

* **To identify the trends and patterns of employee satisfaction.**
* **To identify the primary factors affecting on Employee satisfaction.**
* **To develop a natural language processing (NLP) model to extract keywords from Reviews.**
* **Develop a recommendation system that offers tailored suggestions.**
* **To develop a Multiclass Sentiment Classification model that classifies employees rating as Positive, Negative and Neutral.**

**Chapter-4**

**Methodology**

**Data Collection:** The Capgemini Employees Review Dataset is collected from the Kaggle.com.

**Data Pre-processing:** An exploratory data analysis (EDA) was conducted to gain meaningful insights from the datasets. This included:

* Understanding the Dataset: Examining the structure, size, and data types of the features.
* Handling Missing Data: Identify the missing values in the dataset.

**Data Visualization:** After cleaning the data, it is important to visualize the data. This step can be done by using Python libraries such as Matplotlib, Seaborn, Plotly, and other libraries.

**Recommendation System:** An employee suggestion system is a formalized process that encourages employees to share ideas, propose improvements, and identify areas for change within an organization. It can be implemented through physical suggestion boxes or digital platforms, and its core principles are inclusivity, transparency, and responsiveness. It ensures diverse perspectives are considered and valued, fosters trust, and reinforces the organization's commitment to continuous improvement.

**Multi-class sentiment classification model:**

Sentiment analysis uses machine learning algorithms like Multinomial Naive Bayes (MNB), Multinomial Logistic Regression, and Long-short-Term Memory (LSTM). MNB classifies text into predefined sentiment categories, while LSTM is versatile for multi-class classification tasks. LSTM captures long-range dependencies in sequential data, making it ideal for sentiment analysis and other natural language processing tasks.

**Model Comparison:**

Based on the accuracy of the three models, compare the performance of the multinomial logistic regression model, the multinomial naive bayes model, and the long-short-term memory (LSTM).

**Conclusion:**

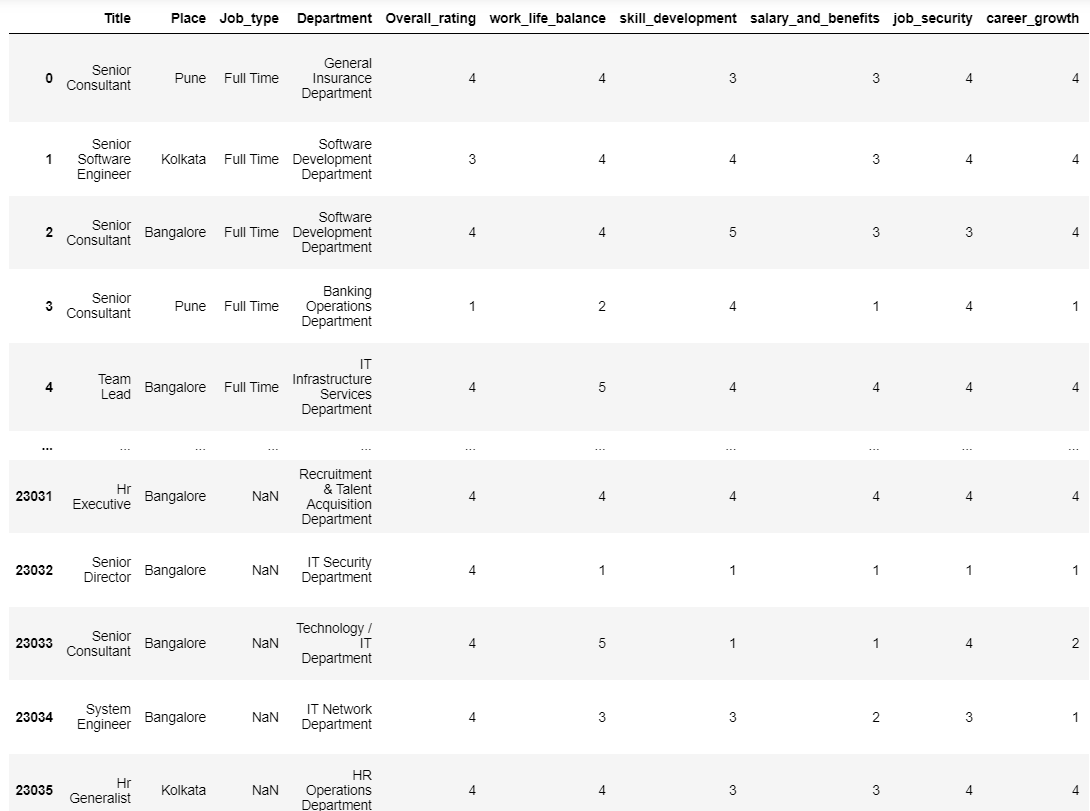
Out of the three models, the LSTM model had the highest accuracy. Thus, it can be observed that for sentiment analysis, the LSTM model is better than both the Naive Bayes and Multinomial Logistic Regression models by only a small margin.

**4.1 Data Description: -**

Data is taken from the Kaggle website. This is large dataset contains 23035 rows and 13 columns.

Here is short overview of Data

* Source: Kaggle.com
* Link: https://www.kaggle.com/datasets/manishkr1754/capgemini-employee-reviews-dataset
* Capgemini Employees Review Dataset



There are some important variables in dataset are:

* **Title:** The job title or part of the hand furnishing the review.
* **Place:** The geographical position or megacity where the hand works.
* **Department:** The specific department or functional area within the association.
* **Overall ratings:** A numerical standing given by the hand for their overall job satisfaction.
* **Work Life Balance:** Rating indicating the work- life balance endured by the hand.
* **Skill Development:** Rating reflecting the openings for skill improvement and growth.
* **Salary and Benefits:** Standing assessing the satisfaction with compensation and benefits.
* **Job Security:** Standing expressing the hand's sense of job security.
* **Career Growth:** Rating indicating the perceived career advancement openings.
* **Work Satisfaction:** Standing showcasing the hand’s pleasure with their work.
* **Likes:** Positive aspects and pros stressed by the hand in their review.
* **Dislikes:** Negative aspects and cons mentioned by the hand in their review.

**4.2 Data Pre-processing:**

Data preprocessing is playing a crucial role in time series analysis. It involves tasks like handling missing values, removing outliers, scaling features, and encoding categorical variables. By preprocessing the data, we ensure that data is in right format and ready for training our models. Data preprocessing is a primary step in machine learning where we prepare our data before training our models. It involves various techniques to clean, transform, and enhance the quality of our data.

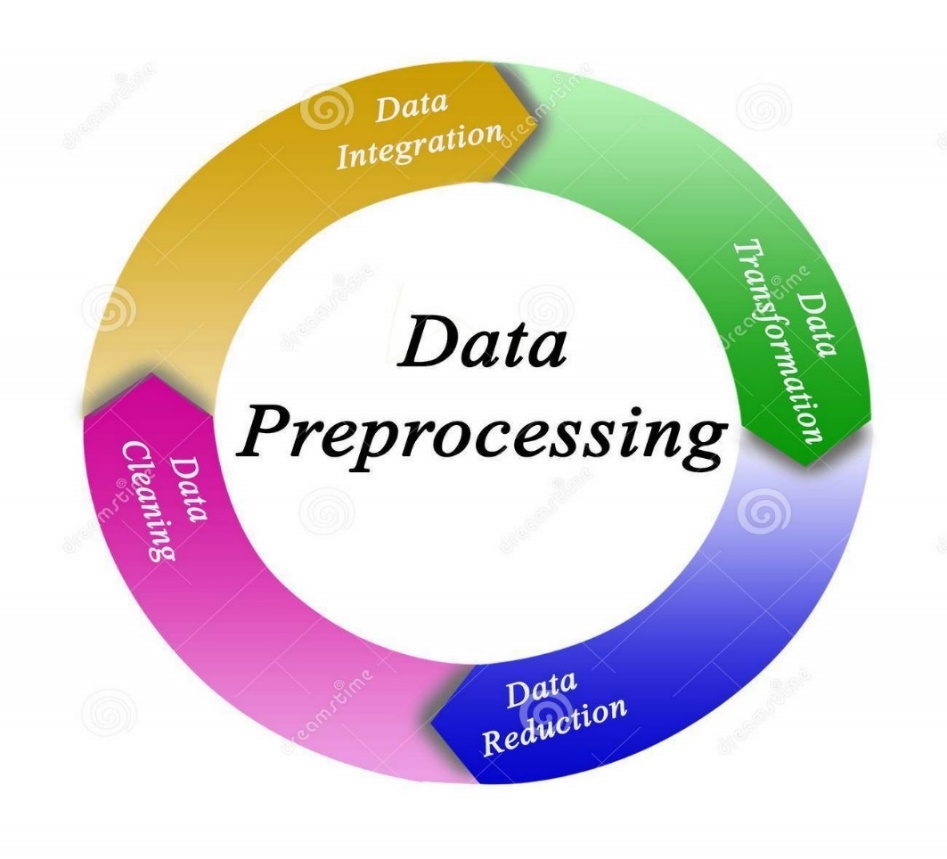
Some common steps in data preprocessing includes:

**1. Data Cleaning.**

**2. Data Integration.**

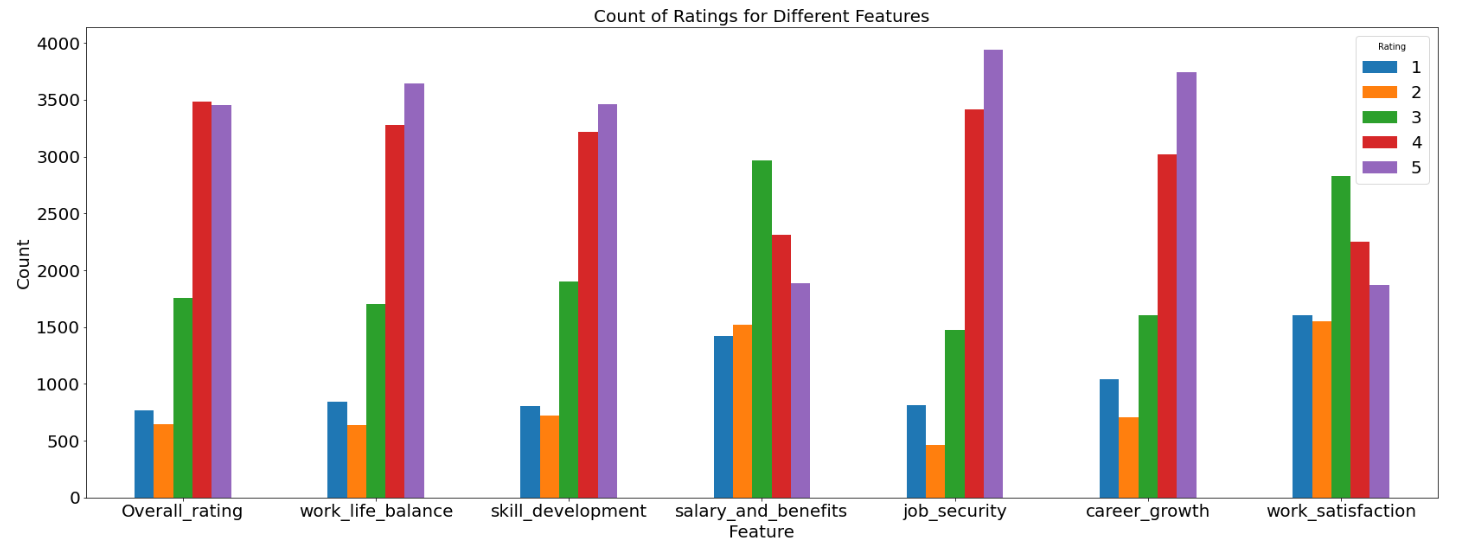
**3. Data Transformation.**

**4. Data Reduction.**



**4.3 Data Visualization:**

1. **Count Plot of ratings for different features**

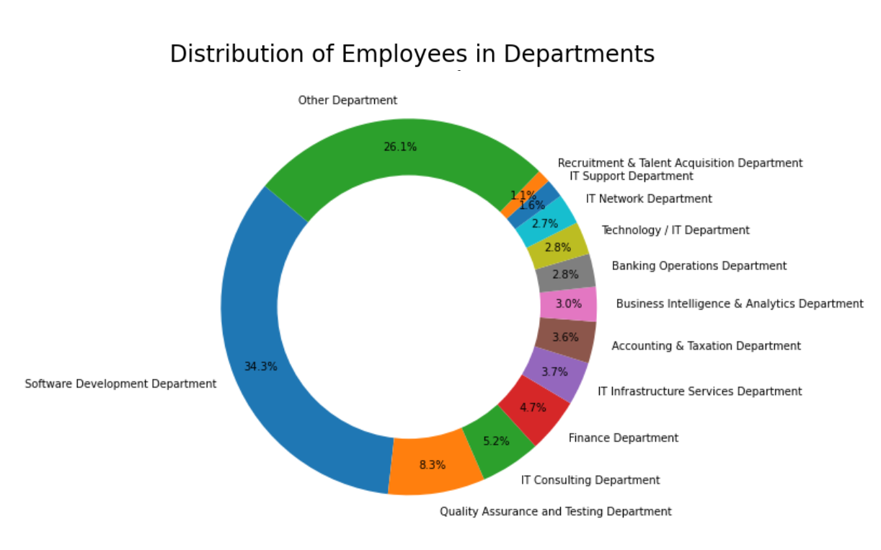
****

**Fig.1**

**Interpretation:**

* In work life balance, job security and career growth the employees give mostly positive rating that means they are satisfied with the company with these features.
* The salary benefits and work satisfaction contain the mostly neutral counts of the employees.

1. **Donut Plot of Employees in department**

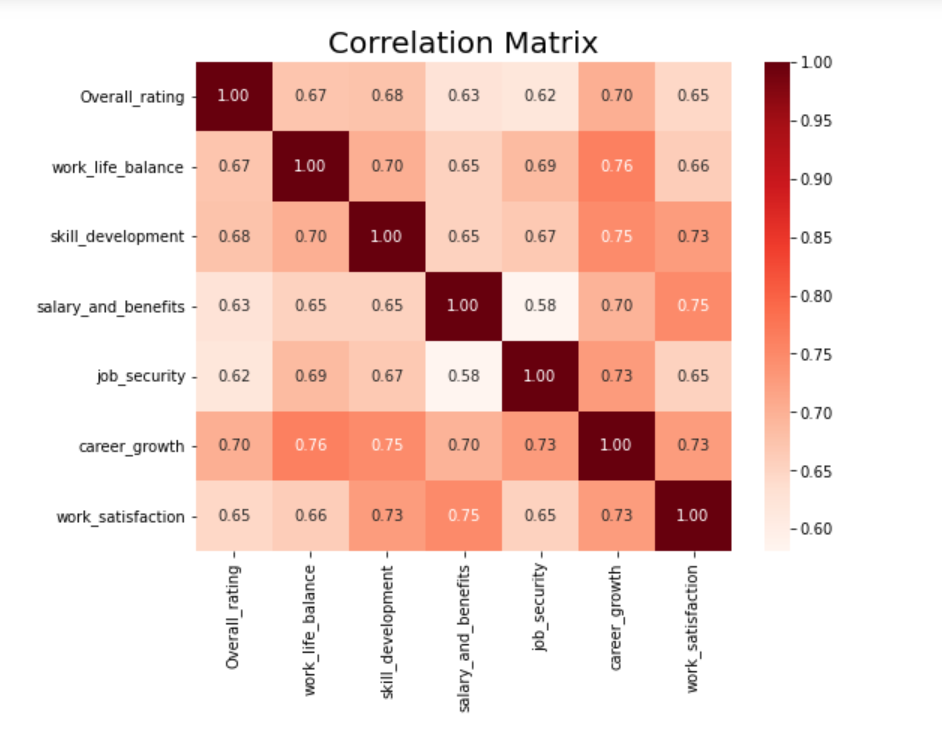
****

**Fig.2**

**Interpretation:**

* In this pie chart, the software development department has the most employee count that is 34.3%.
* The 26.1% employee in the Capgemini company work in the other departments like, Telecom department, Marketing department, cyber security, resource management etc.

1. **Correlation Heatmap**

****

**Fig.3**

**Interpretation:**

* The strong correlation in between career growth and work life balance is (0.76).
* The weakest correlation between salary and benefits and job security (0.58).
* This means that employees who are satisfied with their jobs also satisfied with other aspects of their work such as work life balance, skill development, salary and benefits.

1. **Word Cloud**

* **Word Cloud for Likes.**

****

**Fig.4**

* **Word Cloud for Dislikes.**

****

**Fig.5**

**4.4 Analysis Using Software**

**Natural Language Processing:**

The ability of a computer program to comprehend spoken and written human language isknown as natural language processing, or nlp. It's part of the artificial intelligence (ai) system.With its origins in the study of languages, nlp has been around for further than 50 times. It hasnumerous practical uses in a range of industries, such as business intelligence, search engines,and medical research.

**What is the process of natural language processing?**

Computers can comprehend natural language just like people thanks to nlp. Natural language processing uses artificial intelligence to process and interpret real-world input spoken or written in a way that a computer can comprehend. Computers have programs to read and microphones to record audio, just as humans have various sensors like ears to hear and eyes to see. Computers have programs to process the inputs that they receive, just as humans have brains to process information. The input is eventually changed into computer-understandable code during processing.

The two primary stages of natural language processing are as follows:

1. Data preprocessing.

2. Algorithm development.

'Preparing' and 'cleaning' text data so that computers can analyze it is known as data preprocessing. Preprocessing identifies textual elements that an algorithm can use and puts data into a format that is practical.

This can be fulfilled in a number of ways, including:

* Tokenization: This is when textbook is broken down into lower units to work with.
* Stop word junking: This is when common words are removed from textbook so unique words that offer the most information about the textbook remain.
* Lemmatization and stemming: This is when words are reduced to their root forms to process.
* Part-of-speech tagging: this is the marking of words according to the part of speech they belong to, such as adjectives, verbs, and nouns.

An algorithm is created to process the data after it has undergone preprocessing. Although there are numerous natural language processing algorithms, but two main types are commonly used:

* Rules-based system: This system makes use of meticulously crafted linguistic rules. This method is still in use today, having been employed early in the development of natural language processing.
* Machine learning-based system: Statistical techniques are used by machine learning algorithms. They are fed training data to help them learn how to perform tasks, and as more data is processed, they modify their approaches accordingly. Natural language processing algorithms use a combination of neural networks, deep learning, and machine learning to refine their own rules via repeated processing and learning.
* **Extract keywords from Reviews by using NLP technique:**

Extracting keywords from employee’s review (Likes and Dislikes) using Natural Language Processing (NLP) techniques is a valuable approach for recommendation systems.

Firstly, we are applied Term Frequency-Inverse Document Frequency (TF-IDF) score NLP techniques to our dataset for extraction of keywords from the employee’s review (Likes and Dislikes).

TF-IDF measures the importance of a term within a document relative to its overall presence in the corpus (collection of documents). It assigns higher weights to terms that appear frequently in a specific document while penalizing those that are common across all document.

Here is a general an overview of the process:

1. Tokenization: Split each Likes and Dislikes into individual words or tokens.

2. Remove Stop words: Remove common words that don't carry much meaning (e.g., "the," "and," "is").

3. Calculate Term Frequency (TF): Count the number of times each word appears in employee’s review (Likes and Dislikes).

4. Calculate IDF (Inverse Document Frequency): IDF is a measure of how important a word is across all employee’s review (Likes and Dislikes). It's calculated as the logarithm of the total number of descriptions divided by the number of descriptions containing the word.

5. Calculate TF-IDF: Multiply TF by IDF for each word in each employee’s review (Likes and Dislikes).

6. Select Keywords: Choose the top TF-IDF scored words as keywords for each employee’s review (Likes and Dislikes).

Here is an overview simplified example in Python using the scikit-learn library:

**Output:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Description** | **Keywords** | **TfIdf\_Scores** |
| 22193 | NothingPolicy is not good | Work | 3.260481 |
| 8947 | This Organization Conduct good learning sessio... | Good | 1.822455 |
| 4547 | Salary on time.\nMNC Work culture\nWork life b... | Is | 1.536754 |
| 3700 | Good to work under this companyi didn't face a... | Security | 0.89385 |
| 508 | As a fresher it is good companyFor experienced... | Job | 1.133523 |
| 6697 | CultureAppraisal\nWork culture\nJob security | Salary | 1.036641 |
| 11109 | Salary appraisalNA. There is nothing which I d... | Like | 1.073958 |
| 4061 | Nothing I like.No job security here plus bad e... | Not | 1.380125 |
| 20427 | Good work place and they follow the ethicsEmpl... | No | 1.536754 |
| 2226 | Job security and work life balancePoor hikes, ... | Life | 3.276802 |

**Table No.1**

**Interpretation:**

From above output, we can conclude that NLP is technique used to extract keywords from text by identifying the most important and relevant words and phrases based on TF-IDF scores.

**Recommendation System:**

An employee suggestion system, often referred to as a suggestion box or idea management system, is a formalized process that encourages employees to contribute their insights, propose improvements, and highlight areas for change within the organization. These systems can take various forms, ranging from traditional physical suggestion boxes to sophisticated digital platforms integrated into company intranets or specialized software solutions.

The core principles of an effective employee suggestion system revolve around inclusivity, transparency, and responsiveness. By soliciting input from all levels of the organization, regardless of job title or department, these systems ensure that diverse perspectives are considered and valued. Moreover, transparent communication regarding the evaluation and implementation of suggestions cultivates trust and reinforces the organization's commitment to continuous improvement.

Key features of successful employee suggestion systems include:

**Accessibility and Ease of Use:** The system should be easily accessible to all employees and simple to navigate, whether through online portals, mobile applications, or physical suggestion boxes strategically placed in the workplace.

**Clear Guidelines and Expectations:** Providing clear guidelines on the types of suggestions welcomed, the submission process, and the criteria for evaluation helps employees understand how best to contribute and increases the likelihood of actionable proposals.

**Timely Feedback and Acknowledgment:** Prompt acknowledgment of submitted suggestions, along with transparent feedback on their evaluation status, demonstrates respect for employees' contributions and encourages ongoing participation.

**Recognition and Rewards:** Recognizing employees for their valuable contributions, whether through monetary incentives, awards, or public acknowledgment, reinforces a culture of appreciation and motivates further engagement.

**Integration with Organizational Goals:** Aligning the suggestion system with the organization's strategic objectives ensures that proposed ideas are relevant, impactful, and conducive to driving desired outcomes.

**Continuous Improvement:** Regularly evaluating the effectiveness of the suggestion system, soliciting feedback from participants, and making iterative improvements based on insights gained ensures its relevance and sustainability over time.

* **Develop a recommendation system that offers tailored suggestions.**

Building employees suggestion system based on department and place which can suggest the nature of the company.

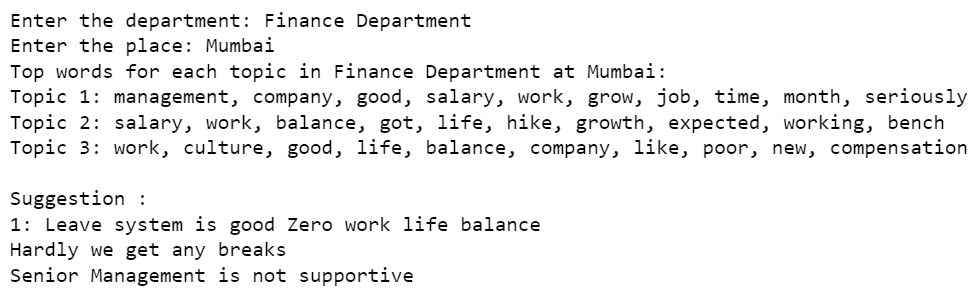
Here's a simplified example using Python with some commonly used libraries:

INPUT:

**Suggestion = Department: Finance Department**

**Place: Mumbai**

OUTPUT:

****

**Sentiment Analysis:**

Sentiment analysis is a technique used to identify thoughts or feelings expressed in a document, such as a review, comment, or social media post. It involves analyzing words, phrases and content in the text to understand whether the opinion is positive, negative or neutral. This analysis is particularly useful for extracting insights from large amounts of textual data, such as social media posts, customer reviews, and ratings or news articles.

Sentiment analysis, also term as view mining or emotional AI, is the field of Natural Language Processing that deals with the identification, extraction, analysis, and understanding of the information in text or speech. It aims to extract and quantify the emotional tone of a piece of text, identifying specific emotions (happy, sad, angry, etc.).

There are several different methods for sentiment analysis. About of the most mutual methods are:

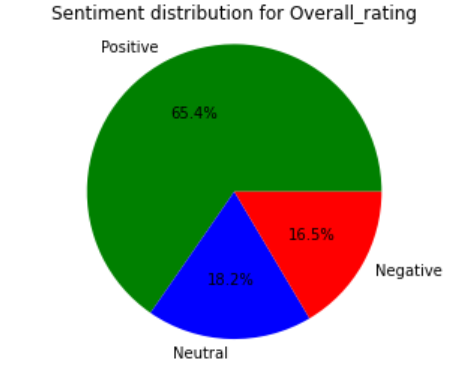
* Lexicon-based methods: These techniques employ a predetermined list of terms and expressions that are connected to either a positive, negative, or neutral emotion. The sentiment of a text is determined by the frequency of these words and phrases.
* Machine learning methods: This technique uses machine learning algorithms to learn the relationship between words and sentiment. The algorithms are trained on a large amount of labelled text, and then they are used to predict the sentiment of new text.
* Hybrid methods: This system combines dictionary-based methods and machine learning methods. This may contribute to raising the sentiment analysis's accuracy.
* Sentiment analysis has become increasingly sophisticated with the advancement of NLP techniques and the availability of large labelled datasets. It is now widely used in various industries to gain valuable insights from the large amount of unstructured text data generated daily.
* **To develop a multi-class sentiment classification model.**

Firstly, we are performing sentiment classification on ratings then calculate the average sentiment score for each rating and we are distributing the ratings column into three category such as positive, negative and neutral as shown in the following table:

|  |  |
| --- | --- |
| Ratings | Sentiment |
| 1 to 2 | Negative Sentiment |
| 3 | Neutral Sentiment |
| 4 to 5 | Positive Sentiment |

**Table No.2**

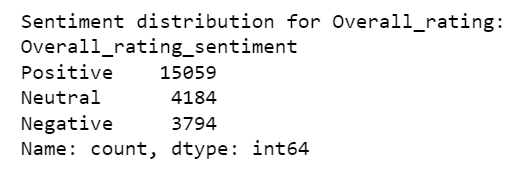
Based on this sentiment classification we draw the pie diagram for visualisation of sentiments and total count of sentiments.



**Interpretation:**

From the above pie diagram of sentiment distribution, we can easily analyse that the count of positive sentiment is highest as compare to neutral and negative sentiment.

Here we are calculating the count of each sentiment category in the column:

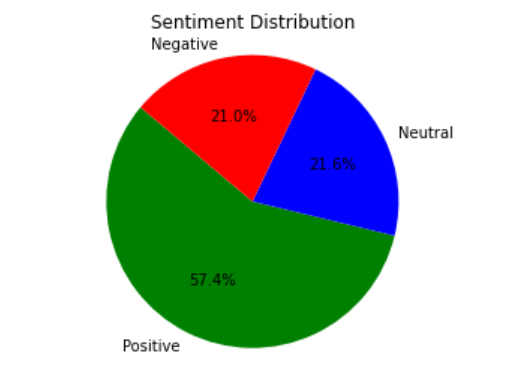


So, by observing the count of sentiments we can say that our data is imbalance so we have to handle the imbalance dataset by using the technique such as

Sampling Techniques:

1. **Under-sampling:** To match the size of the minority class, this method entails reducing the number of data points from the majority class. Random under-sampling is a simple method that randomly removes instances from the majority class. However, it may discard valuable information and reduce the overall size of the training set.
2. **Over-sampling:** This approach involves replicating data points from the minority class to increase its representation in the training set. Oversampling at random only results in duplicates of the minority class. However, it can lead to overfitting and introduce artificial correlations.
3. **Synthetic Minority Oversampling Technique (SMOTE):** SMOTE is an effective oversampling technique that creates new synthetic minority class instances by interpolating between existing minority class data points. This approach helps preserve the inherent distribution of the original data while increasing the size of the minority class.
4. **Use ensemble methods:** such as bagging and boosting, with algorithms that inherently handle imbalanced data well. Algorithms like Random Forest and AdaBoost often perform better on imbalanced datasets.

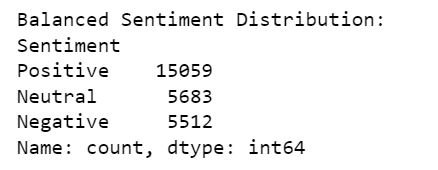
For our dataset we are implement the Random Over-Sampling technique to handle the imbalance sentiment class present in the data. So, by using the Over-sampling method it reduces number of neutral sentiments up to number of negative sentiments and positive sentiment class remain same as shown in above pie diagram



**Interpretation:**

From the above bar diagram of sentiment distribution, we can easily analyse that the count of positive sentiment and negative sentiment are same & neutral sentiment remains as it is.

Here we are calculating the count of each sentiment category in the column:



After handling the imbalance data then we have to develop a multi-class sentiment classification model that classifies ratings as positive, negative, or neutral.

* **Sentiment analysis using multi-class sentiment classification model.**

1. **Multinomial logistic regression:** Multinomial Logistic Regression is a versatile algorithm used for multi-class classification tasks. It extends logistic regression to handle multiple target classes by employing a SoftMax function, making it suitable for sentiment analysis and other text classification tasks.

**Output:**

|  |  |
| --- | --- |
| Accuracy | 0.769948 |

**Classification Report:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column1 | Precision | recall | F1-score | Support |
| Negative | **0.77** | **0.76** | **0.76** | **1120** |
| Neutral | **0.56** | **0.35** | **0.43** | **1113** |
| Positive | **0.81** | **0.93** | **0.87** | **3018** |
| Accuracy |  |  | **0.77** | **5251** |
| Macro avg | **0.71** | **0.78** | **0.69** | **5251** |
| Weighted avg | **0.75** | **0.77** | **0.75** | **5251** |

**Table No.3**

1. **Multinomial Naive Bayes:** Multinomial Naive Bayes (MNB) is a simple and effective machine learning algorithm commonly used for sentiment analysis. It is based on Bayes' theorem, which provides a framework for calculating probabilities based on conditional statements. In the context of sentiment analysis, MNB is used to classify text into predefined sentiment categories, such as positive, negative, or neutral.

**Output:**

|  |  |
| --- | --- |
| Accuracy | 0.785183 |

**Classification Report:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column1 | Precision | recall | F1-score | Support |
| Negative | **0.76** | **0.78** | **0.77** | **1120** |
| Neutral | **0.56** | **0.60** | **0.58** | **1113** |
| Positive | **0.89** | **0.86** | **0.87** | **3018** |
| Accuracy |  |  | **0.79** | **5251** |
| Macro avg | **0.74** | **0.74** | **0.74** | **5251** |
| Weighted avg | **0.79** | **0.79** | **0.79** | **5251** |

**Table No.4**

1. **LSTM (Long-short term memory):** LSTM is an important intermittent neural network (RNN) armature designed to overcome the evaporating grade problem. With specialized memory cells, it effectively captures long-range dependencies in sequential data like text, making it ideal for sentiment analysis and other natural language processing tasks.

**Output:**

|  |  |
| --- | --- |
| Accuracy | 0.790516 |

**Classification Report:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column1 | Precision | recall | F1-score | Support |
| Negative | **0.78** | **0.77** | **0.77** | **1120** |
| Neutral | **0.59** | **0.54** | **0.56** | **1113** |
| Positive | **0.86** | **0.89** | **0.88** | **3018** |
| Accuracy |  |  | **0.79** | **5251** |
| Macro avg | **0.74** | **0.73** | **0.73** | **5251** |
| Weighted avg | **0.79** | **0.79** | **0.79** | **5251** |

**Table No.5**

**Model Evaluation:**

To assess the performance of the trained MNB classifier on the testing set, we calculated the following metrics:

**Accuracy:** The proportion of correctly classified reviews

**Precision:** The proportion of positive predictions that are actually correct

**Recall:** The proportion of actual positive reviews that are correctly identified

=

**F1-score:** The harmonic mean of precision and recall

**Comparison of models:**

1. Multinomial logistic regression
2. Multinomial Naive Bayes
3. LSTM (Long-short term memory)

Compare the performance of Multinomial logistic regression model, Multinomial Naive Bayes model and LSTM (Long-short term memory) on the basis of the Accuracy of three models.

The accuracy of Multinomial logistic regression

* **Accuracy=0.769948**

The accuracy of Multinomial Naive Bayes

* **Accuracy=0.785183**

The accuracy of LSTM

* **Accuracy=0.790516**

**Result and Conclusion:**

LSTM model has the highest accuracy among the three models. Therefore, the conclusion would be that the LSTM model performs slightly better than both the Naive Bayes and Multinomial Logistic Regression models for sentiment analysis.

* **Overall Accuracy:** The overall accuracy of the MNB classifier is 0.79, indicating that it correctly classified 79% of the product ratings. This is a relatively high accuracy, suggesting that the classification is able to effectively distinguishing between positive, negative, and neutral sentiments.
* **Sentiment Class Performance:**

1. **Negative Reviews:** The classifier has a precision of 0.78 for negative reviews, meaning that only 78% of the reviews it classified as negative were actually negative.
2. **Neutral Reviews:** The classifier has a precision of 0.59 and a recall of 0.54 for neutral reviews.
3. **Positive Reviews:** The classifier has a precision of 0.86 and a recall of 0.89 for positive reviews.

**Chapter-5**

**Conclusion and Discussion**

* We can conclude that factors like work-life balance, skill development, salary, benefits, job security, and career growth positively impact employee work satisfaction, leading to increased job satisfaction.
* We can conclude that employees generally are satisfied with work-life balance, job security, and career growth, but neutral ratings for salary benefits and work satisfaction suggest that these areas may require improvement or further attention from the company.
* We can conclude that NLP is technique used to extract keywords from text by identifying the most important and relevant words and phrases based on TF-IDF scores.
* We can Build suggestion system based on department and place which can suggest the nature of the company.
* We can conclude that the overall accuracy of the LSTM is 0.8036, indicating that it correctly classified 80.36% of the employee’s ratings. This is a relatively high accuracy, suggesting that the classification is able to effectively distinguishing between positive, negative, and neutral sentiments.

**Chapter-6**

**Limitation and Future Scope**

**Limitations**

The presence of missing values in the data. In some columns, missing values may not be replaceable due to their nature. This can potentially impact the accuracy and reliability of the Model as missing data may lead to incomplete insights or skewed results.

Natural language processing (NLP) models may struggle with understanding context in employee reviews. This can lead to inaccuracies in keyword extraction, sentiment classification, and recommendation generation.

Data is imbalance, which could impact the performance of the sentiment analysis model. Balancing the data by either oversampling the minority classes or under sampling the majority class can help mitigate this issue, but it may also introduce bias or lead to a loss of valuable information.

The multiclass sentiment classification model's generalization ability may be limited if the training data does not sufficiently cover the diverse range of employee sentiments found across various contexts and organizations. This could result in the model being less effective at accurately classifying sentiments in new or unseen data, especially if the sentiment distribution in the new data differs significantly from the training data.

**Future Scope**

By identifying trends and patterns in employee satisfaction, you can help organizations understand what factors contribute to their employees' overall satisfaction, leading to improved retention rates and productivity.

The NLP model to extract keywords from reviews enables the organization to efficiently analyze large volumes of feedback, gaining deeper insights into specific issues or trends affecting employee satisfaction.

Identifying the primary factors affecting employee satisfaction can provide valuable insights for organizations to focus on areas that need improvement, such as work-life balance, career growth opportunities, and compensation.

The multiclass Sentiment Classification model can help organizations understand the sentiment behind employee ratings, allowing them to address issues and improve employee satisfaction in a targeted manner.

**Chapter-7**

**Reference**

1. Ramadhan, W. P., Novianty, S. A., & Setianingsih, S. C. (2017, September). Sentiment analysis using multinomial logistic regression. In 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC) (pp. 46-49). IEEE.
2. Minaee, S., Azimi, E., & Abdolrashidi, A. (2019). Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models. arXiv preprint arXiv:1904.04206.
3. Parveen, H., & Pandey, S. (2016, July). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. In 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT) (pp. 416-419). IEEE.
4. Hasanli, H., & Rustamov, S. (2019, October). Sentiment analysis of Azerbaijani twits using logistic regression, Naive Bayes and SVM. In 2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-7). Ieee.
5. <https://www.kaggle.com/datasets/manishkr1754/capgemini-employee-reviews-dataset>