Employee Opinion Tracker

Application of

Sentiment Analysis & Topic Modeling

NLP PROJECT BY TEAM A

Indonesia AI
AI for Everyone, AI for Indonesia

Spärck

Meet the Spärck Team

Alfandy Surya

Topic Modeling

Alfandy Surya

Efrad Galio

Deployment: Streamlit

Efrad Galio

Muhammad Habibullah

Sentiment Analysis

Muhammad Habibullah

Alfian Ali Murtadlo

Data Preparation & Preprocessing

Alfian Ali Murtadlo

Anton Pranowo Medianto

Domain Experts & EDA

Anton Pranowo Medianto

Background &
Problem Statement

In today's competitive labor market, retaining top people is critical to company success. However, employee engagement and satisfaction are volatile. Thus result in turnover, lower productivity, and a bad work atmosphere.

Traditional ways of getting employee input, such as annual surveys or exit interviews, are typically insufficient and fail to capture continuing employee mood. Employees may be cautious to submit honest criticism using these approaches, fearing punishments or believing their views will not be heard.

Problem Statement:

Company lack a consistent and continuing method for gathering honest and actionable feedback from employees. Absence of real-time knowledge makes it difficult to identify organization climate and manage employee complaints before they become disengaged or turnover.

Objective & Scope

Objective

Employee Opinion Tracker is a comprehensive aimed at understanding organizational climate through consistent and real time analysis of employee opinions.

By leveraging sentiment analysis and topic modeling techniques, this project provides valuable insights into the workplace environment.

The results are presented through an interactive Streamlit dashboard, making it easy to visualize and interpret employee opinion. This allows companies to:

- Gain real-time insight into employee mood.
- Identify and handle employees' problems based on topics before escalate.
- Make data-driven decisions to create a more positive and productive workplace.

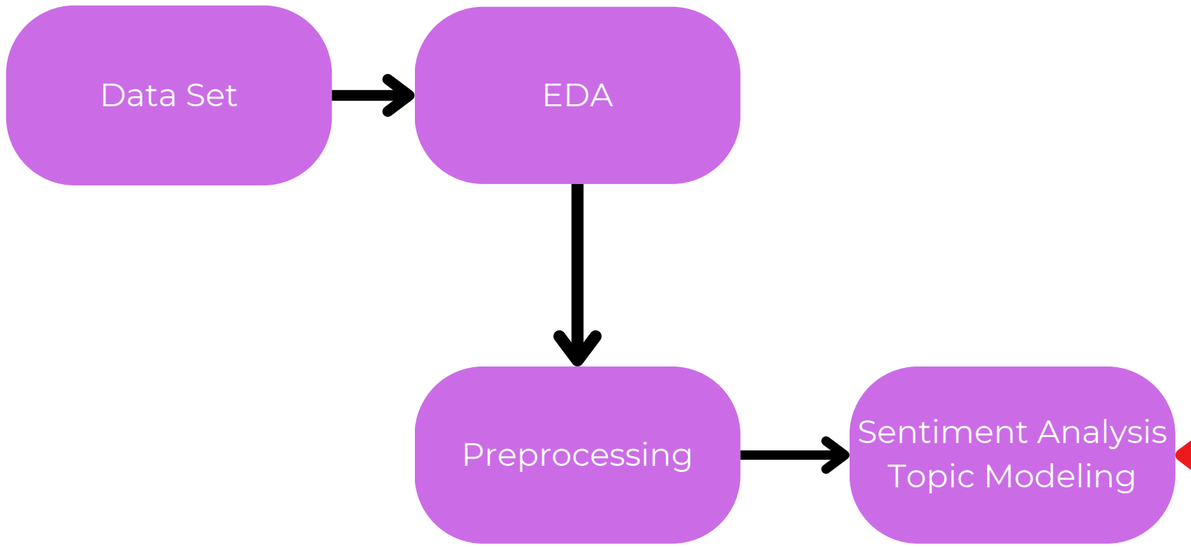
Scope

Scope of this project is limited:

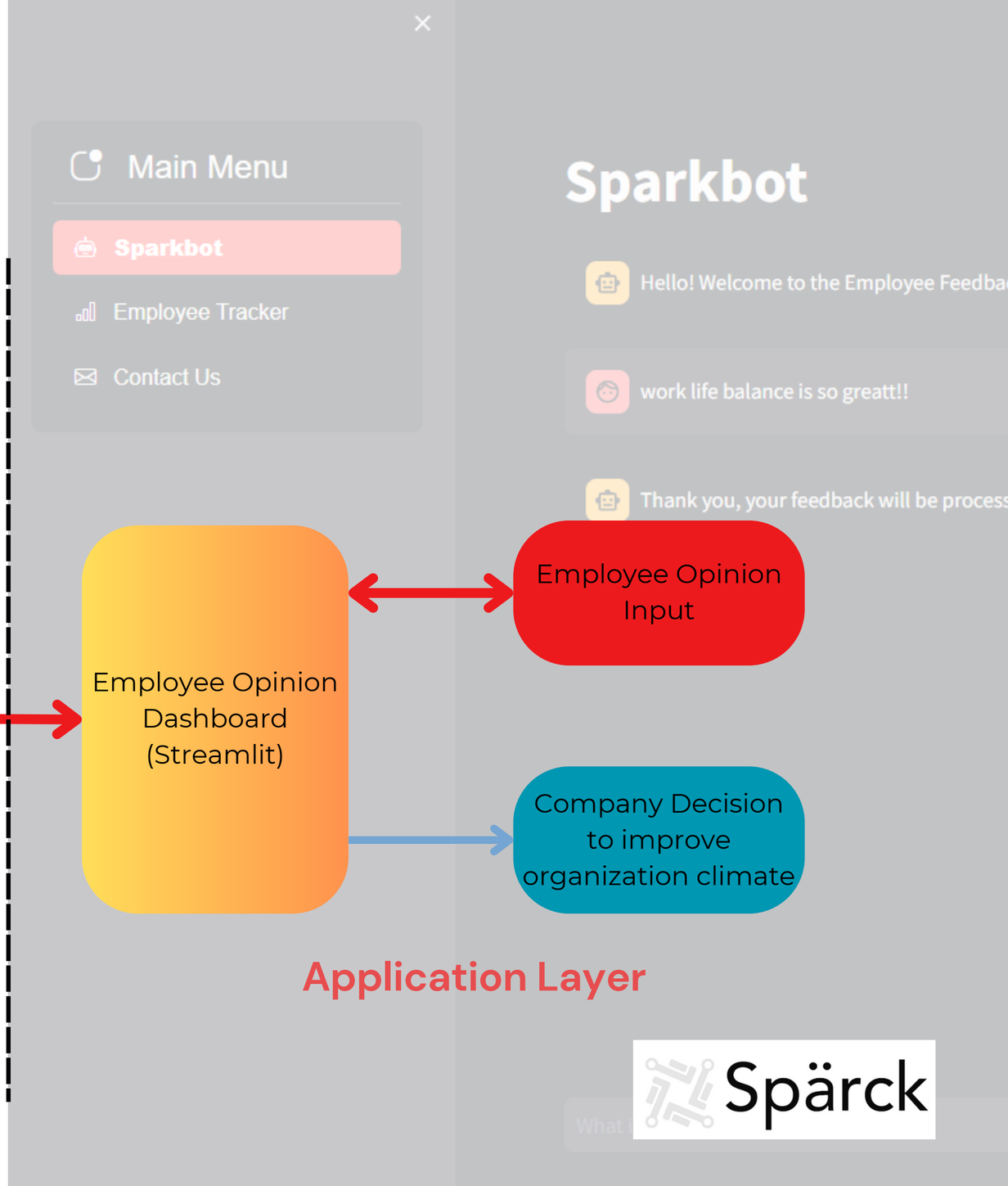
- DatasetsCapgemini_Employee_Reviews_from_AmbitionBox
- Sentiment analysis with BERT
- Topic modeling using LDA, LLM and XGBoost
- Deployment with Streamlit

Spärck

Flow



Development Layer



Application Layer

Spärck

Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27013 entries, 0 to 27012
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Title                25932 non-null  object
 1   Place                24613 non-null  object
 2   Job_type             11576 non-null  object
 3   Department           22103 non-null  object
 4   Date                 25935 non-null  object
 5   Overall_rating       25918 non-null  float64
 6   work_life_balance    26997 non-null  float64
 7   skill_development    26996 non-null  float64
 8   salary_and_benefits  26967 non-null  float64
 9   job_security         26963 non-null  float64
 10  career_growth        26951 non-null  float64
 11  work_satisfaction    26929 non-null  float64
 12  Likes                23884 non-null  object
 13  Dislikes             22986 non-null  object
dtypes: float64(7), object(7)
memory usage: 2.9+ MB
```

Null

Title	1081
Place	2400
Job_type	15437
Department	4910
Date	1078
Overall_rating	1095
work_life_balance	16
skill_development	17
salary_and_benefits	46
job_security	50
career_growth	62
work_satisfaction	84
Likes	3129
Dislikes	4027

Duplicatel

Number of duplicate entries in Title: 22768
Number of duplicate entries in Place: 26182
Number of duplicate entries in Job_type: 27007
Number of duplicate entries in Department: 26391
Number of duplicate entries in Date: 26273
Number of duplicate entries in Overall_rating: 27007
Number of duplicate entries in work_life_balance: 27007
Number of duplicate entries in skill_development: 27007
Number of duplicate entries in salary_and_benefits: 27007
Number of duplicate entries in job_security: 27007
Number of duplicate entries in career_growth: 27007
Number of duplicate entries in work_satisfaction: 27007
Number of duplicate entries in Likes: 9514
Number of duplicate entries in Dislikes: 9464

Dataset
the employee satisfaction result
Capgemini_Employee_Reviews_from_AmbitionBox
27.013 Entries

14 Attributes

- **4 Attributes Information** (Title, Place Job_Type, Department, Date)
- **7 Attributes Satisfaction Scale of topics** (Overall, Work Life Balance, Skill Development, Salary and Benefit, Job Security, Career Growth, Work Satisfaction)
- **2 Attributes employee opinion** (Likes and Dislikes)

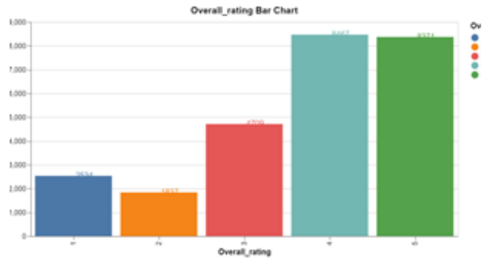
Summary

- Lots of null data and duplicate data
- Information & satisfaction scale data duplicate due to selection template
- Likes & Dislikes, duplicate and null due to no response or comment

Spärck

EDA

EDA provides an overview of the puzzle with the next process in the selection of topics to be used



Most Frequent Word Analysis

N Gram Analysis

Co Occurrence Analysis

Result satisfaction scale

- **Overall is Good**
  - Good : Work Life Balance, Skill Development, Career Growth, Job Security, Work Satisfaction
  - Moderate :Salary & Benefit Moderate

Result related to topic

- Dislikes : Work, Salary, management, appraisal, growth, balance, compensation
- Likes : Culture, Balance, Security, Environment, Team, Opportunity, Salary

Result related to topic

- Likes & Dislikes : Work life balance, Work Environment, working culture, security, career growth

Result related to topic

- Likes & Dislikes : Work Life Balance, Salary, Career Growth, Environment, Salary, Appraisal, Culture

Spärck

Preprocessing

Text Cleaning Process:

1. Lower casing
2. Remove Stop words
3. Remove Punctuation
4. Remove special character
5. Processing Slang Words
6. Remove Extra Spaces
7. Processing number

Text Normalization Process:

8. Lemmatization + Finalization

Spärck

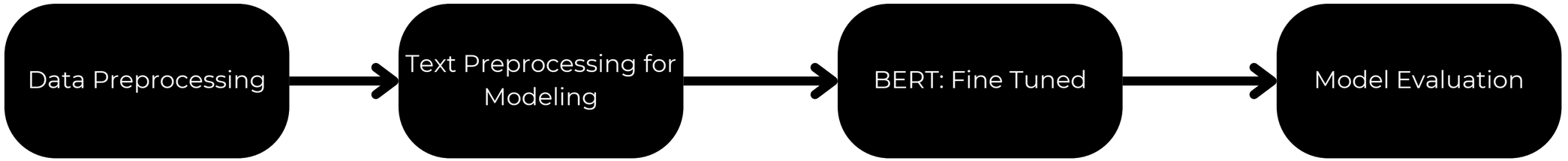
Preprocessing Example (Before & After)

Text Preprocessing Stage	OUTPUT
LIKES	Deserved candidates are promoted promptly. Unbiased in providing opportunities to employees, regardless of their gender or any other thing
LIKES_CLEANED	deserved candidate promoted promptly unbiased providing opportunity employee regardless gender thing
DISLIKES	Culture, micro management, unprofessional behavior, lack of sensitivity, pathetic HR and PMO
DISLIKES_CLEANED	culture micro management unprofessional behavior lack sensitivity pathetic hr pmo

Sentiment Analysis

Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral. In this case, sentiment is divided into two, positive and negative. We fine tuned the Bert Base Uncased pretrained model. The following is the data splitting strategy and development flow of employee review topic modeling.

Data Splitting Strategy:

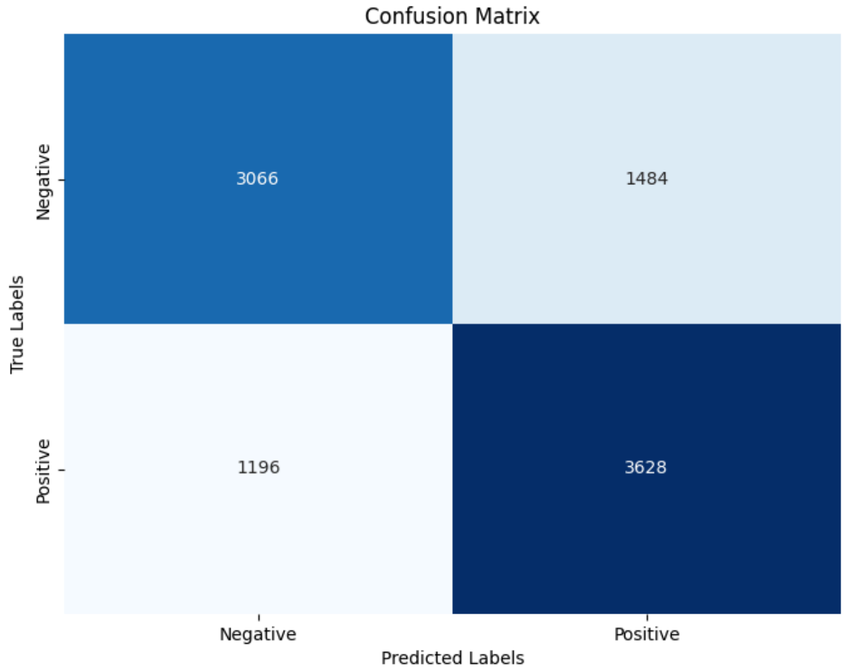


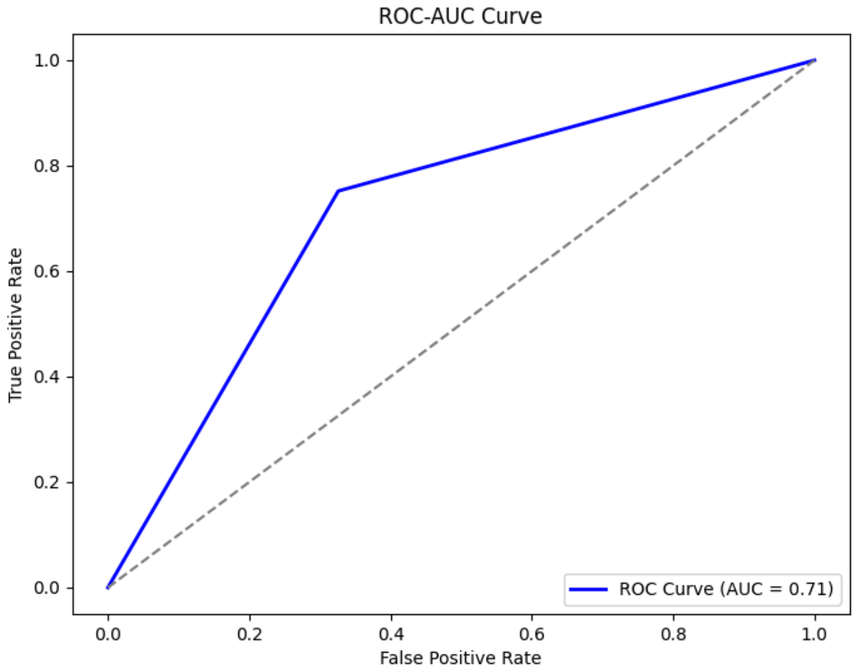
- Remove duplicate data
- Data Splitting + List
- Text Tokenization & Load Model (bert-base-uncased)

Fine tuned:

- epoch: 3
- bath size: 32
- learning rate: 3e-5
- logging steps: 50

Modeling and Evaluation (Sentiment) - Result





Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Bert	0.714103	0.709703	0.752073	0.730274	0.71296

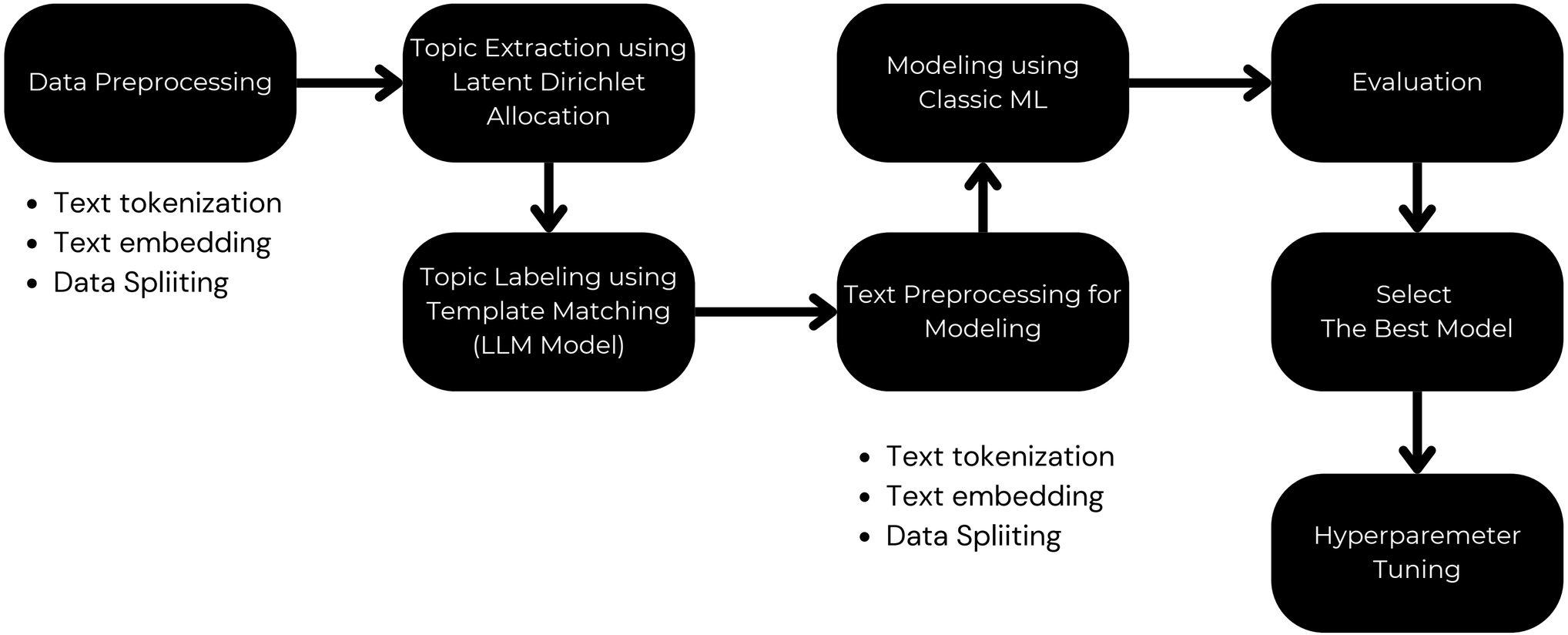
Overall the model has better performance so we choose BERT FIne Tuned for sentiment analysist task.

Sentiment Analysis (Inference)

Input	OUTPUT
BEST COOPERATIVE SUPPORTING COLLEGUES WORK LIFE BALANCE GOOD LEARNING OPPORTUNITY	Positive
WASTE COMPANY WORK HIKE TECHNOLOGY UPGRADE THEY CARE DRESS EMPLOYEE GIVING STAR RATING ALSO WASTE	Negative
MAKING EMPLOYEE WAIT ONSITE TRAVEL CHEAP COMPENSATION	Negative
REALLY GOOD OPPORTUNITY WORK CAPGEMINI	Positive

Topic Modeling

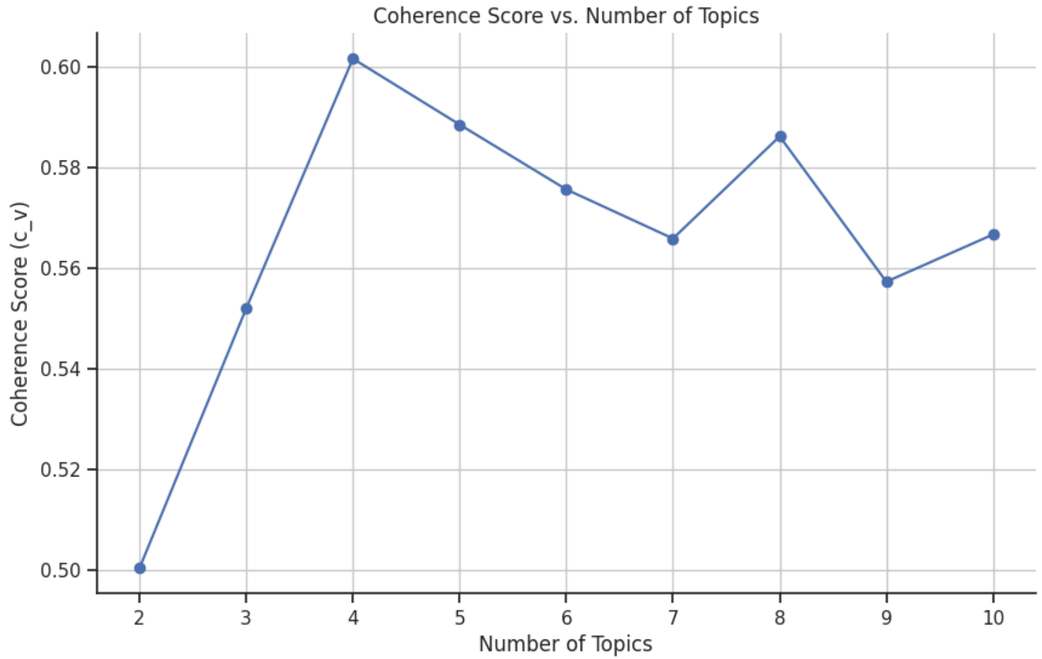
We also predict the topic of the review using topic modeling. Topic modeling is a type of statistical modeling used in natural language processing and text mining to discover the abstract "topics" that occur within a collection of documents. This is the development flow of employee review topic modeling.



Topic Extraction using Latent Dirichlet Allocation

LDA is one of the most popular topic modeling algorithms. It assumes that each document is a mixture of a small number of topics and that each word in the document is attributable to one of the document's topics. Here is the result of LDA

Determine the optimal number of topics using coherence vs no. of topic plot:



Coherence is used to evaluate the quality of the topics produced by the topic modeling algorithm.

LDA result:

- (0, '0.177*"work" + 0.090*"life" + 0.083*"balance" + 0.053*"growth"')
- (1, '0.030*"job" + 0.025*"company" + 0.024*"security" + 0.023*"employee"')
- (2, '0.069*"salary" + 0.053*"hike" + 0.042*"appraisal" + 0.035*"compensation"')
- (3, '0.211*"good" + 0.108*"work" + 0.070*"culture" + 0.057*"company"')

Potential topics:

- Work-Life Balance, Professional Development and Growth (1)
- Employment Environment, Organizational Stability, and Job Security (2)
- Salary, Performance Appraisals, and Compensation (3)
- Company or Organizational Culture (4)
- Other Topics (to accommodate unknown topics) (0)

Topic Labeling using Template Matching

We use a pre-trained sentence-transformers model for data labeling: all-MiniLM-L6-v2 (https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2). Basically, we use the predefined topic to make sure that the topic is consistent.

Template matching process:

```
def find_most_relevant_topic(text, topic_embeddings, predefined_topics):
    text_embedding = model.encode(text)
    similarities = util.pytorch_cos_sim(text_embedding, topic_embeddings)
    most_relevant_index = similarities.argmax()
    return predefined_topics[most_relevant_index]
```

- Encoding the Text: Convert the input text to an embedding vector.
- Calculating Similarities: Compute cosine similarities between the text embedding and each of the topic embeddings.
- Identifying the Most Relevant Topic: Find the index of the highest similarity score.
- Returning the Topic: Return the topic name corresponding to the highest similarity score.

Sample result:

```
Text 1:
text: deserved candidate promoted promptly unbiased providing opportunity employee regardless gender thing
main topic: Salary, Performance Appraisals, and Compensation
--------------------------------------------------------------------------------------------------------------
Text 2:
text: designation promotion good salary increment also required
main topic: Salary, Performance Appraisals, and Compensation
--------------------------------------------------------------------------------------------------------------
Text 3:
text: got lot learning platform monthly learning plan well people encourage cloud certificating
main topic: Other Topics
--------------------------------------------------------------------------------------------------------------
Text 4:
text: get fully tech project variable pay no appraisal worked hard one recognise you increment max max
main topic: Salary, Performance Appraisals, and Compensation
```

Label proportion:

```
Topic
Other Topics                                                    0.279407
Work-Life Balance, Professional Development and Growth          0.230155
Company or Organizational Culture                               0.180751
Salary, Performance Appraisals, and Compensation                0.179416
Employment Environment, Organizational Stability, and Job Security    0.130271
Name: proportion, dtype: float64
```

Modeling and Evaluation (Topic Modeling)

Data Splitting Strategy:

60% train	20% validation	20% test
27.860 rows	9.287 rows	9.287 rows

use 3 fold cross validation

Experimentation:

Word2Vec SkipGram	Word2Vec CBOW
• XGBoost • LGBM • Random Forest • Logistic Regression • etc.	• XGBoost • LGBM • Random Forest • Logistic Regression • etc.

Modeling and Evaluation (Topic Modeling) - Result

Top 5 Model (Word2Vec SkipGram) – 3 Fold CV

	Model	Accuracy	AUC	Recall	Prec.	F1
xgboost	Extreme Gradient Boosting	0.8034	0.9636	0.8034	0.8037	0.8029
lightgbm	Light Gradient Boosting Machine	0.8012	0.9634	0.8012	0.8015	0.8004
rf	Random Forest Classifier	0.7905	0.9578	0.7905	0.7920	0.7897
et	Extra Trees Classifier	0.7897	0.9564	0.7897	0.7925	0.7891
gbc	Gradient Boosting Classifier	0.7834	0.0000	0.7834	0.7838	0.7827

Top 5 Model (Word2Vec CBOW) – 3 Fold CV

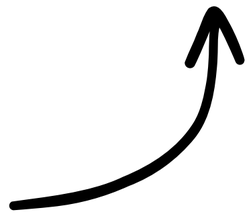
	Model	Accuracy	AUC	Recall	Prec.	F1
xgboost	Extreme Gradient Boosting	0.7950	0.9594	0.7950	0.7954	0.7946
lightgbm	Light Gradient Boosting Machine	0.7945	0.9594	0.7945	0.7949	0.7938
rf	Random Forest Classifier	0.7869	0.9561	0.7869	0.7883	0.7860
et	Extra Trees Classifier	0.7858	0.9548	0.7858	0.7883	0.7852
gbc	Gradient Boosting Classifier	0.7753	0.0000	0.7753	0.7760	0.7747

As you can see SkipGram embedding method is overall better than CBOW. So we select the top 2 best models using SG to compare the performance of test and validation set.

Validation and Test Result

Model Name	Accuracy	AUC	Recall	Precision	F1 Score
XGBoost SG - Val	0.815118	0.966954	0.800002	0.808618	0.803749
LGBM SG - Val	0.808873	0.965483	0.793151	0.803125	0.797453
XGBoost SG - Test	0.819856	0.967633	0.806952	0.816100	0.810932
LGBM SG - Test	0.813072	0.966919	0.798508	0.810376	0.803380

Overall the XGBoost model is better than LGBM so we choose LGBM for Hyperparameter tuning.

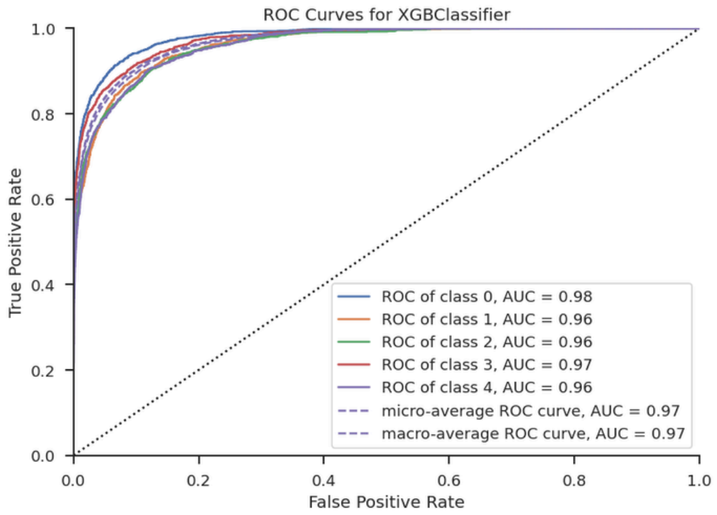


Modeling and Evaluation (Topic Modeling) - Best Model

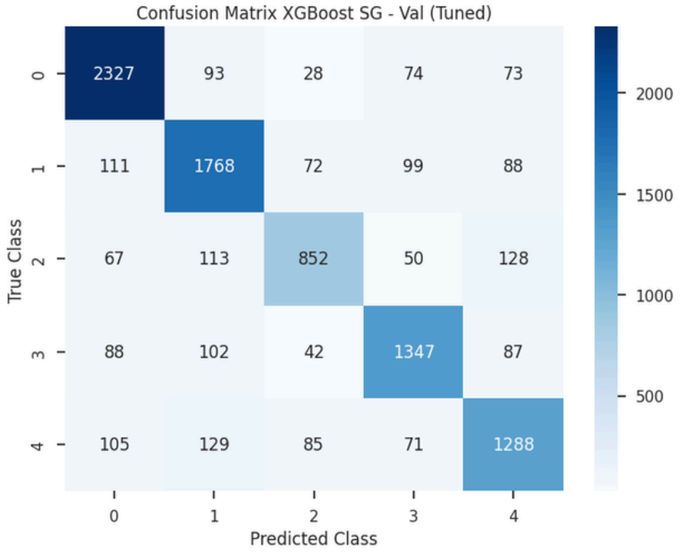
We use optuna to tune the hyperparameters to improve efficiency. Here is the best params and the performance result

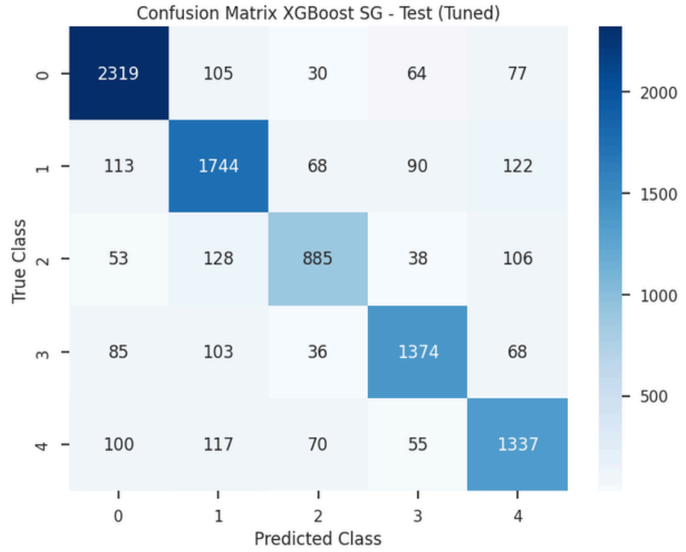
```
best_params = {'n_estimators': [199],
 'max_depth': [27],
 'min_child_weight': [9],
 'gamma': [0.050619310950326166],
 'learning_rate': [0.07836589625857665],
 'subsample': [0.8575260193186514],
 'colsample_bytree': [0.8473654274503567],
 'reg_alpha': [0.5599374574894134],
 'reg_lambda': [0.43310182069056535]}
```

ROC-AUC Curve:



Confusion matrix:





Model Name	Accuracy	AUC	Recall	Precision	F1 Score
XGBoost SG - Val (Tuned)	0.81641	0.968209	0.80078	0.809761	0.804664
XGBoost SG - Test (Tuned)	0.824701	0.969539	0.812436	0.820904	0.816131

Conclusion:

XGBoost using SkipGram the best combination model. The baseline model resulting 81.5% for validation and 81.9% for test. The tuned model resulting 81.6% for validation and 82.4% for test.

Results: Steamlit Dashboard

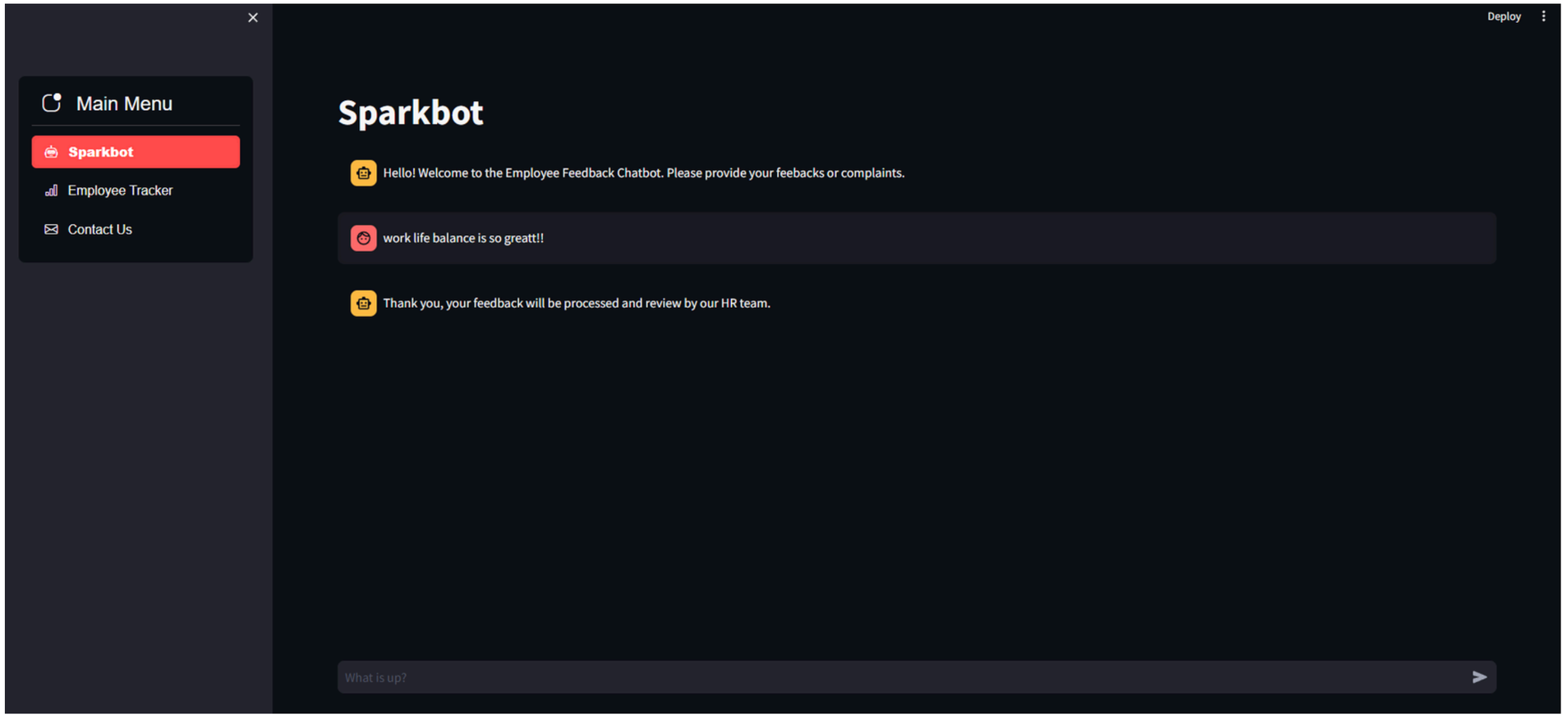


Figure 1. An Employee giving feedbacks in SparkBot

Results: Steamlit Dashboard

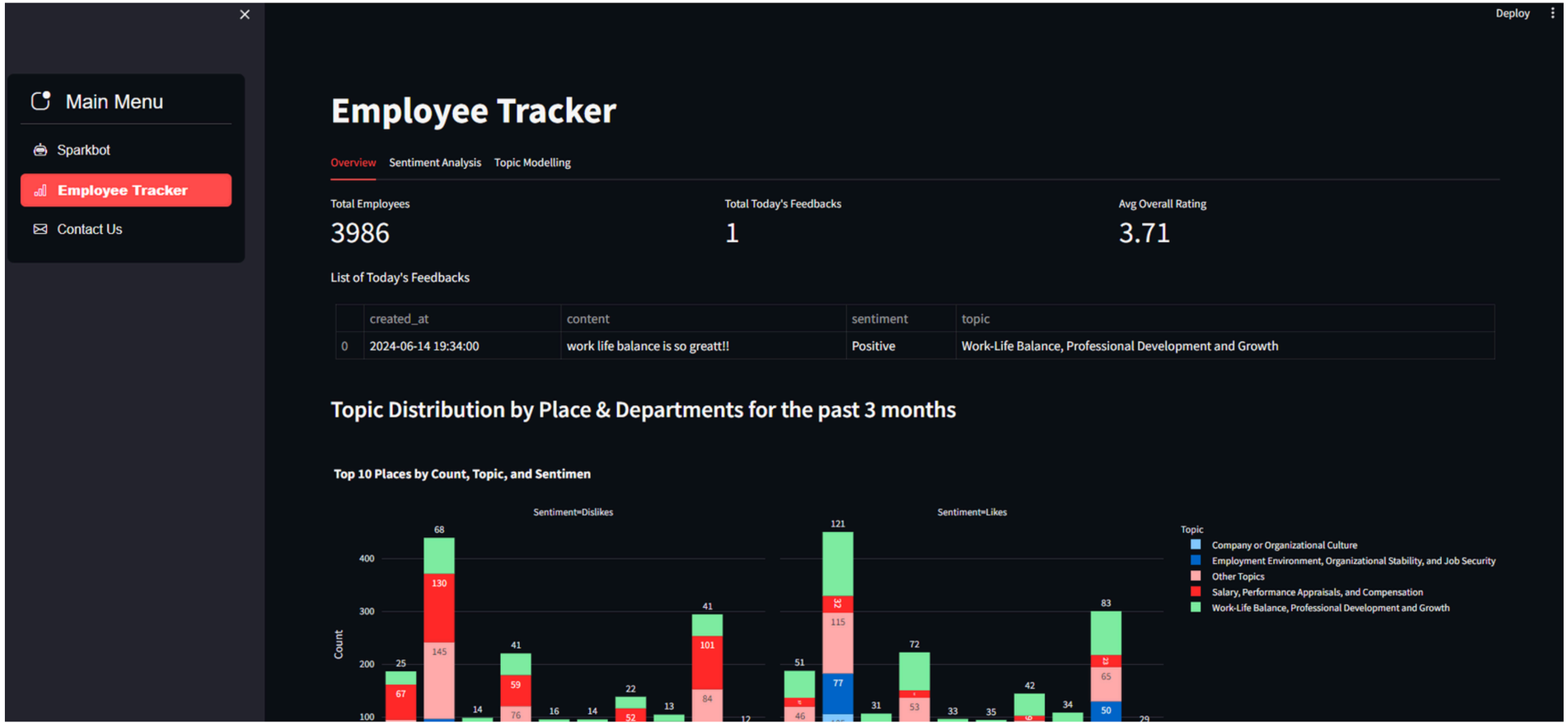


Figure 2. Employee feedbacks will be shown in Employee Tracker

Real-world Application

An employee opinion tracker is a dashboard that allows Company HR Division to collect and evaluate real time feedback from their employees to understand patterns over time and find opportunities for improvement organization climate and working environment.

Therefore employee opinion tracker, help HR to :

- Improve Employee engagement when employees believe their opinions are heard.
- Assist in identifying workplace difficulties, such as poor morale or communication concerns.
- Better decision-making: Understanding employee sentiment allows firms to make better judgments about rules, processes, and culture.

Future Improvement

- Try to use a dataset in Bahasa Indonesia
- Fine-Tuned Sentiment model using more larger data
- Add more feature in Streamlit Dashboard

Conclusion

Within this project we've successfully:

- Create a quite accurate Sentiment Analysis Model using LLM (BERT).
- Be able to categorize topic from Employee Input Feedbacks using Topic Modeling Model.
- Build a Streamlit Dashboard for tracking Employee Feedbacks.

Contact Us

Don’t hesitate to contact us for further inquiries or any collaborations.

Alfandy Surya

Alfandy Surya

Alfian Ali Murtadlo

Alfian Ali Murtadlo

Anton Pranowo Medianto

Anton Pranowo Medianto

Efrad Galio

Efrad Galio

Muhammad Habibullah

Muhammad Habibullah