IIT BHILAI

MACHINE LEARNING

Company360: A Deep Learning Approach for a complete analysis of companies

Final Report

Submitted to: Dr. Gagan Raj Gupta

Associate Professor

Team Members:

Moyank Giri	12310830	M.Tech DSAI
Aishika Nandi	12140120	B.Tech CSE

Contents

List of Figures

1 Introduction and Problem Motivation

In recent years, there has been a significant shift in the priorities of working professionals while evaluating the potential employers. Factors such as work-life balance, organizational culture and values, diversity and inclusion, job security have become increasingly critical in addition to salary compensation. The challenge lies in bridging the information gap, providing candidates with comprehensive insights into both the advantages and drawbacks of specific companies which enables them to make informed and strategic career decisions that promote long-term stability and job satisfaction.

This document details a project which addresses this need by creating a one-stop solution that caters to the uncertainties and lack of awareness that candidates often face when evaluating potential employers. We aim to achieve this through aspect-based analysis and information summarization of the abundant information available online. This valuable data includes employee reviews sourced from various job portals, content from online platforms, news articles and transcripts from videos where employees provide insights into their day-to-day work experiences.

2 Proposed solution approach

The proposed solution includes majorly 3 modules:

- Company Review Analysis
- Company Reputation Analysis
- Employee Work-Life Analysis

The details are as shown in Figure 1

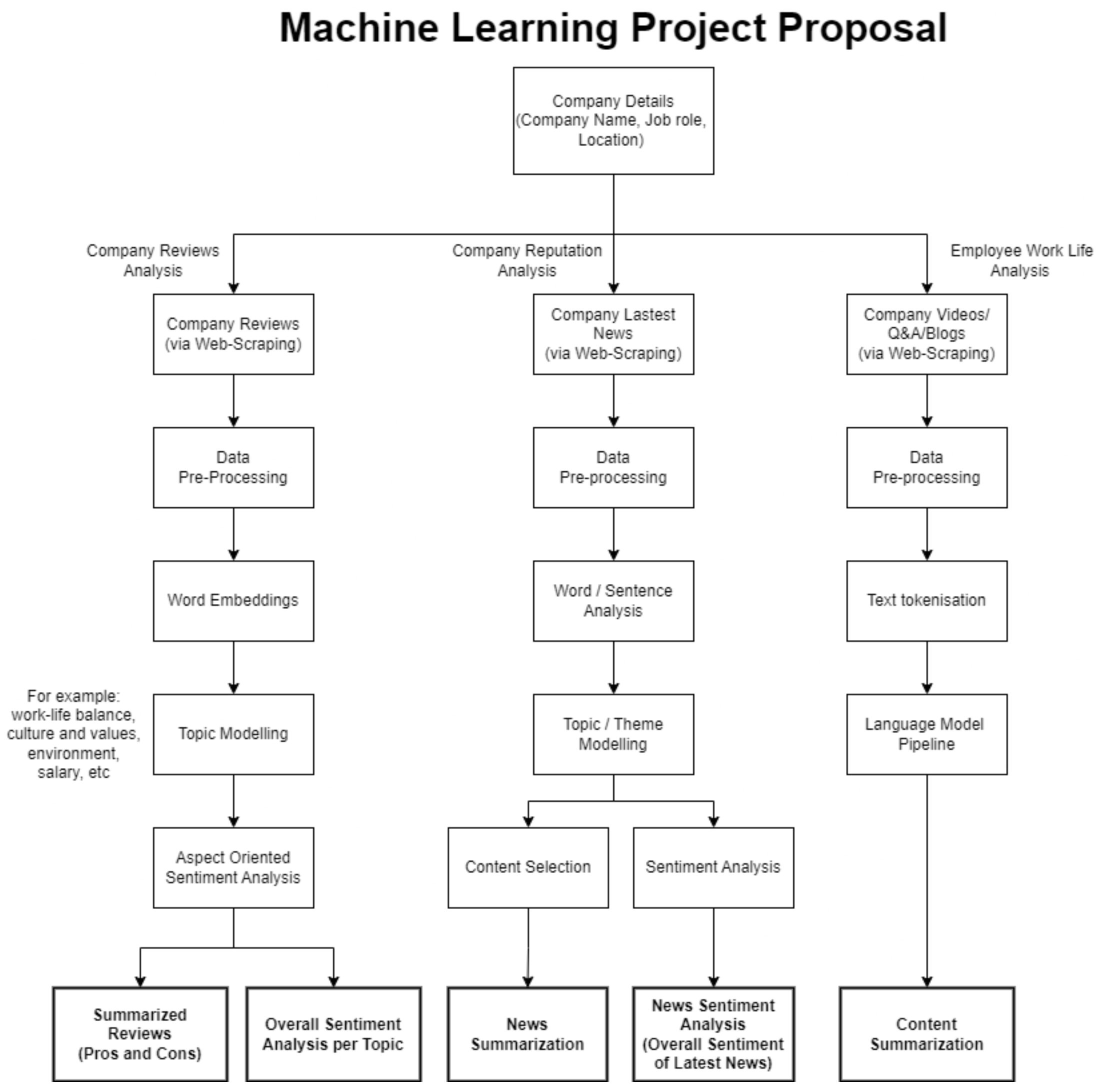

Machine Learning Project Proposal
Company Details (Company Name, Job role, Location)
Company Reviews Analysis
Company Reputation Analysis
Employee Work Life Analysis
Company Reviews (via Web-Scraping)
Company Lastest News (via Web-Scraping)
Company Videos/ Q&A/Blogs (via Web-Scraping)
Data Pre-Processing
Data Pre-processing
Data Pre-processing
Word Embeddings
Word / Sentence Analysis
Text tokenisation
For example: work-life balance, culture and values, environment, salary, etc
Topic Modelling
Topic / Theme Modelling
Language Model Pipeline
Aspect Oriented Sentiment Analysis
Content Selection
Sentiment Analysis
Summarized Reviews (Pros and Cons)
Overall Sentiment Analysis per Topic
News Summarization
News Sentiment Analysis (Overall Sentiment of Latest News)
Content Summarization

Figure 1: High Level Design Diagram

As shown in Figure 1 the 3 major modules take input as company details such as Company Name, Job Role, Location etc. Using these details each modules makes use web-scraping and/or web-crawling methods to extract the necessary data from various renowned websites such as Glassdoor (For company reviews), NDTV, Twitter (For News), Linkedin, Youtube, Quora (For Work-life Analysis). This collected data is then pre-processed using various pre-processing methods employed in Natural Language Processing (NLP) such as stopword removal, Tokenization, Punctuation Mark Removal etc. This pre-processed data will then be converted to Word Embeddings and will be analysed. These word-embeddings will then be passed to for Topic Modelling / Language Processing for Aspect Based Sentiment Analysis and Summarization.
Finally, the result obtained from the model would be

- Summarized reviews about company and Overall sentiment across topics
- Summarized Recent News and Overall Sentiment of the news

- Additional Summary of Company Analysis via Videos/Blogs/Q&A

3 Implemented Solution

3.1 Data Collection

Various data sources were used for scraping information out of the sites. The details are as follows

- Company Reviews Analysis
  - Web Scraping employee blogs from renowned sites like ”Glassdoor” (Shown in Code)
  - Aspect Based Sentiment Analysis (ABSA) Dataset [Link].
  - Google News Dataset [Link].
  - Company Reviews from employers [Link]
  - Company Reviews from Blind App [Link]
- Company Reputation Analysis
  - Web Scraping for news from Reputed sites such as MoneyControl, CrunchBase (Shown in Code)
  - Extreme Summarization Dataset [Link]
- Employee Work-life Analysis
  - Web Scraping employee blogs from renowned sites like ”Quora” (Shown in Code)
  - Amazon Fine Food Reviews Dataset[Link].
  - Multi-News Dataset [Link]

3.2 Pre-processing

The pre-processing has been applied to the collected data. Pre-processing majorly includes some commonly used techniques such as Punctuation Removal, Stop Word Removal etc and these are detailed below:

- Company Reviews Analysis
  - Punctuation Removal, Stop Word Removal, Stemming, Lemmatization
- Company Reputation Analysis
  - HTML Extraction, Data cleaning, Prefixing, Truncation (if needed)
- Employee Work-Life Analysis
  - HTML Extraction, Stop word removal, special characters removal, white spaces removal, Contraction mapping, Remove any text inside the parenthesis, plural removal.

3.3 Model Selection

Here, for each module multiple models were explored which included Machine Learning models, Deep Learning Models and Transformer Models. These are as detailed below:

- Company Reviews Analysis
  - The models are trained and tested on the existing reviews collected from open-source data repositories such as Github.
  - Aspect Based Sentiment Analysis majorly had 2 parts: Aspect Extraction and Aspect Sentiment Analysis.
  - For both parts Machine Learning and **Transformer Models** were used
  - The machine learning models trained are **Naive Bayes Classifier, Support Vector Machine and Random Forest Classifier**
- Company Reputation Analysis
  - For news summarization we have trained/ applied 2 models namely **GPT2 Model and Transformers**
  - The transformer model is fine-tuned on ”Extreme Summarization” Dataset and tested using the RougeL metric
- Employee Work-Life Analysis
  - We opted for a sequence-to-sequence model with attention mechanisms, specifically an **LSTM-based encoder-decoder architecture.**
  - This was chosen to capture contextual dependencies and understand the relationships between words in the input text.
  - We have also fine tuned a pre-trained model: **The T5 Transformer model** due to its versatility in handling various natural language processing tasks, including summarization of large articles

3.4 Feature Engineering

In this part, we explored multiple feature-extraction methods which are currently been used in the field of NLP. The details are as follows:

- Company Review Analysis
  - For the Aspect Based Sentiment Analysis, Feature Engineering is done using tagging methods for Named-Entity Recognition such as POS (Part-Of-Speech) tagging (or) BIO (Begin, Interior, Outside) Tagging for Machine Learning Models
  - For Transformer Models, we make use of Tokenizer provided along with the Transformer Model such as AutoTokenizer
- Company Reputation Analysis
  - For both Transformer Models, we make use of Tokenizer provided in the HuggingFace library such as AutoTokenizer

- Employee Work-Life Analysis
  - Tokenization is performed with Keras' Tokenizer on training data for converting text to integer sequences.
  - Padded sequences ensure uniform length, and vocabulary size is calculated from the tokenizer's word index.

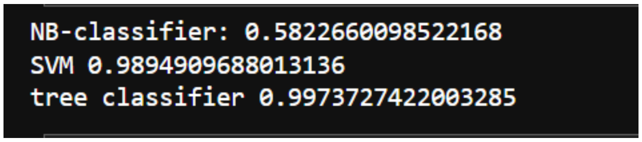
3.5 Hyper-parameter Tuning

Hyper-parameter tuning was done in ordeer to check and verify performance and tune out the best performing model. The details are as follows:

- Company Reviews Analysis
  - We have done hyperparameter tuning for Decision Tree Model, where we tested between a single decision tree and an ensemble of decision trees and we observed that even a single decision tree was performing well when compared to a ensemble (Random Forest).
- Company Reputation Analysis
  - Various Learning Rates were tested and the best learning rate was chosen at lr = 2e-5
- Employee Work-Life Analysis
  - We have conducted hyperparameter tuning on latent dimensions (hidden units) in the encoder and decoder layers.
  - The model explores different architectures and finds that with 500 hidden neurons (latent dimensions), the model fetched the least error.

3.6 Results

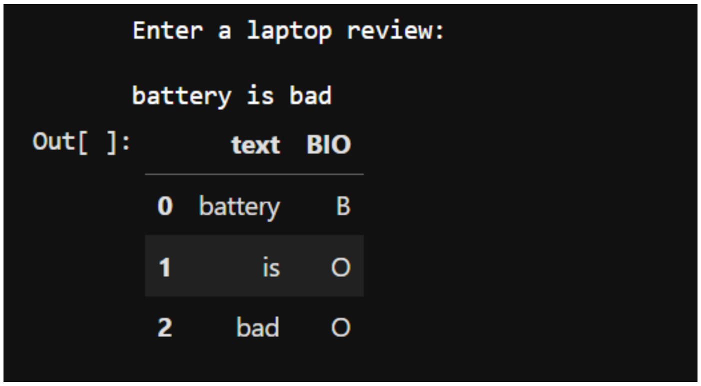
- Company Reviews Analysis
  - The first module was the "Aspect Extraction" module where we made use of 3 machine learning modules namely Naive Bayes Classifier, SVM and Decision Tree Classifier. The accuracies obtained for Aspect Extraction are as shown in Figure 5



```
NB-classifier: 0.5822660098522168
SVM 0.9894909688013136
tree classifier 0.9973727422003285
```

Figure 2: Accuracies Of Models

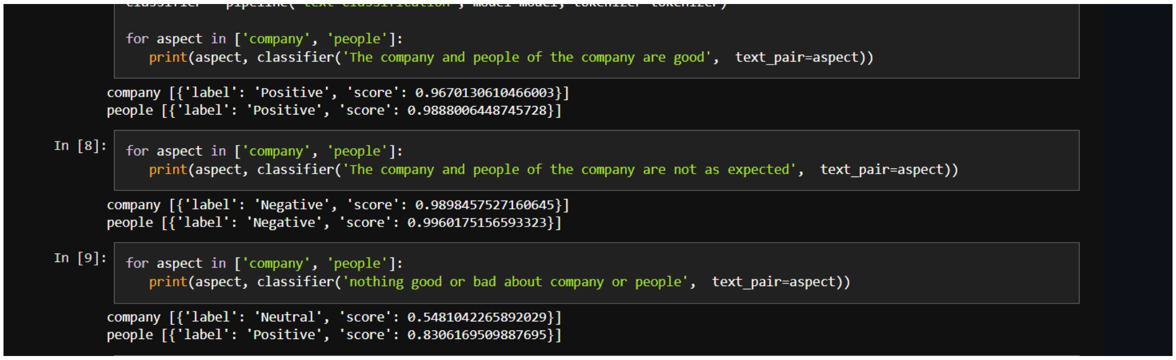
  - The BIO tagging output obtained on custom input is as shown in Figure 3
  - On the other hand, the aspect based sentiment analysis performance was insufficient. Naive Bayes Accuracy was 37%, Decision Tree was 51.7% and SVC was 47.6%



```
Enter a laptop review:

battery is bad
Out[ ]:        text  BIO
          0  battery   B
          1       is   O
          2      bad   O
```

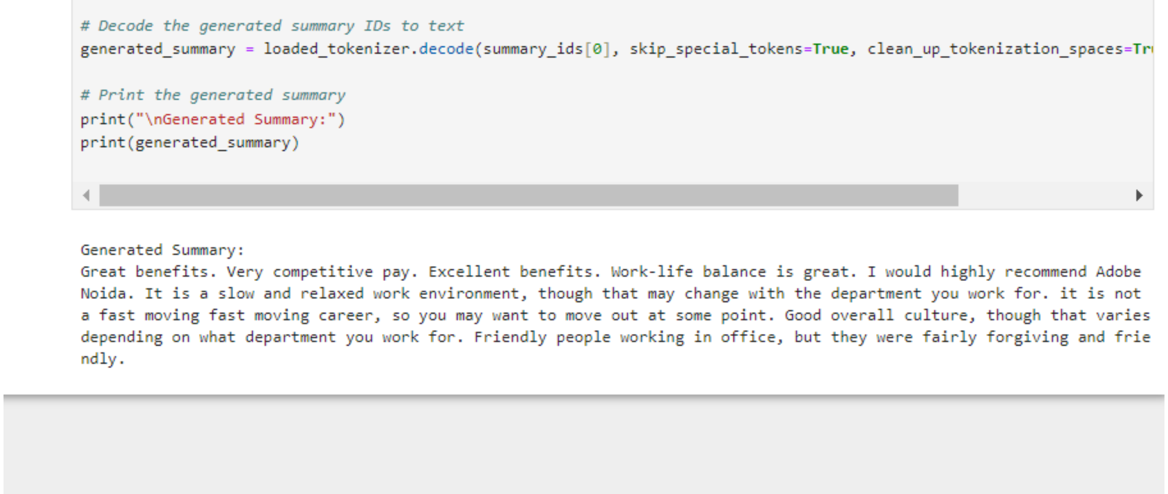
Figure 3: BIO Tagging

– For improved results, we made use of a Transformer model for ABSA. The results are as shown in figure 4



```
         classifier = pipeline('text-classification', model=model, tokenizer=tokenizer)

         for aspect in ['company', 'people']:
            print(aspect, classifier('The company and people of the company are good',  text_pair=aspect))

        company [{'label': 'Positive', 'score': 0.9670130610466003}]
        people [{'label': 'Positive', 'score': 0.9888006448745728}]

In [8]:  for aspect in ['company', 'people']:
            print(aspect, classifier('The company and people of the company are not as expected',  text_pair=aspect))

        company [{'label': 'Negative', 'score': 0.9898457527160645}]
        people [{'label': 'Negative', 'score': 0.9960175156593323}]

In [9]:  for aspect in ['company', 'people']:
            print(aspect, classifier('nothing good or bad about company or people',  text_pair=aspect))

        company [{'label': 'Neutral', 'score': 0.5481042265892029}]
        people [{'label': 'Positive', 'score': 0.8306169509887695}]
```

Figure 4: Reviews Aspect Based Sentiment Analysis



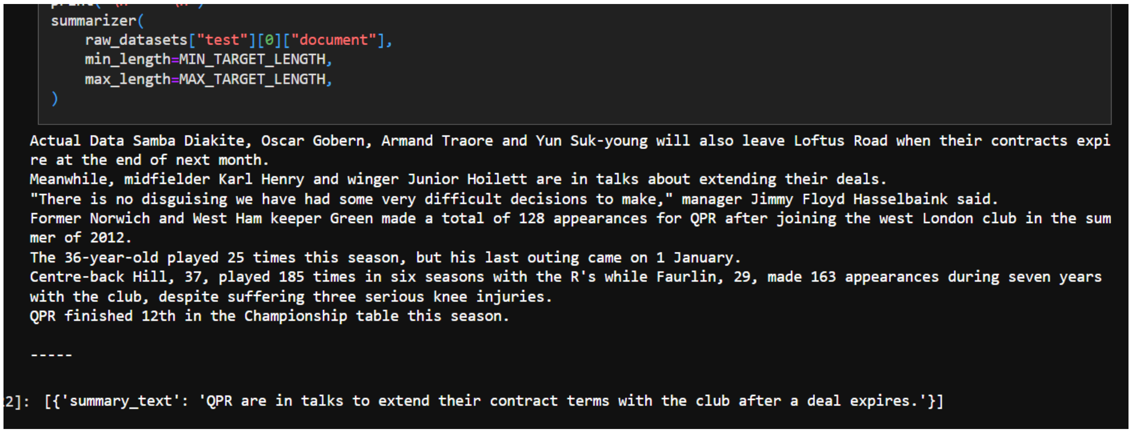
```
# Decode the generated summary IDs to text
generated_summary = loaded_tokenizer.decode(summary_ids[0], skip_special_tokens=True, clean_up_tokenization_spaces=Tr

# Print the generated summary
print("\nGenerated Summary:")
print(generated_summary)
```

```
Generated Summary:
Great benefits. Very competitive pay. Excellent benefits. Work-life balance is great. I would highly recommend Adobe
Noida. It is a slow and relaxed work environment, though that may change with the department you work for. it is not
a fast moving fast moving career, so you may want to move out at some point. Good overall culture, though that varies
depending on what department you work for. Friendly people working in office, but they were fairly forgiving and frie
ndly.
```

Figure 5: Company Reviews Summarization

–

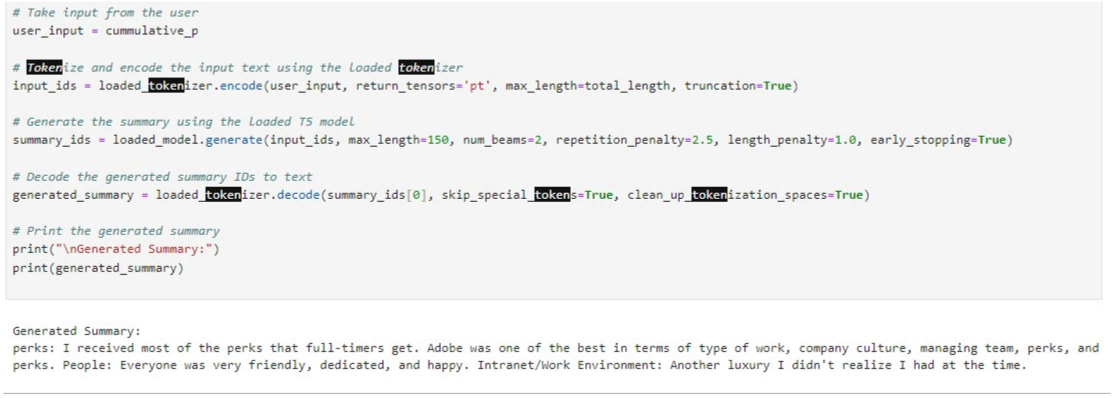
- Company Reputation Analysis
  - The fine-tuned Transformer model used for News Summarization was evaluated using RougeL score where it was able to gain a score RougeL score of 0.2092 in only 2 epochs. Example summarization is as shown in Figure 6:



```
    summarizer(
        raw_datasets["test"][0]["document"],
        min_length=MIN_TARGET_LENGTH,
        max_length=MAX_TARGET_LENGTH,
    )

Actual Data Samba Diakite, Oscar Gobern, Armand Traore and Yun Suk-young will also leave Loftus Road when their contracts expi
re at the end of next month.
Meanwhile, midfielder Karl Henry and winger Junior Hoilett are in talks about extending their deals.
"There is no disguising we have had some very difficult decisions to make," manager Jimmy Floyd Hasselbaink said.
Former Norwich and West Ham keeper Green made a total of 128 appearances for QPR after joining the west London club in the sum
mer of 2012.
The 36-year-old played 25 times this season, but his last outing came on 1 January.
Centre-back Hill, 37, played 185 times in six seasons with the R's while Faurlin, 29, made 163 appearances during seven years
with the club, despite suffering three serious knee injuries.
QPR finished 12th in the Championship table this season.

-----

2]:  [{'summary_text': 'QPR are in talks to extend their contract terms with the club after a deal expires.'}]
```

Figure 6: News Summarization

- Employee Work-Life Analysis
  - In the trained model, the loss decreased from 3.25 to 2.47 during fine-tuning over four epochs, demonstrating improved summarization performance.
  - The pre-trained model achieved an initial loss of 5.86, gradually reducing to 1.67 after the first epoch itself, indicating effective adaptation to the new dataset.
  - The pre-trained model outperformed the trained model as it was trained over a diverse and large dataset whereas we could train the LSTM model upto a certain extent due to lack of extensive computational resources. Nevertheless, both the models provide insights into the actual text that is to be summarised. An example summarization can be seen in Figure 7



```
# Take input from the user
user_input = cummulative_p

# Tokenize and encode the input text using the loaded tokenizer
input_ids = loaded_tokenizer.encode(user_input, return_tensors='pt', max_length=total_length, truncation=True)

# Generate the summary using the loaded T5 model
summary_ids = loaded_model.generate(input_ids, max_length=150, num_beams=2, repetition_penalty=2.5, length_penalty=1.0, early_stopping=True)

# Decode the generated summary IDs to text
generated_summary = loaded_tokenizer.decode(summary_ids[0], skip_special_tokens=True, clean_up_tokenization_spaces=True)

# Print the generated summary
print("\nGenerated Summary:")
print(generated_summary)
```

```
Generated Summary:
perks: I received most of the perks that full-timers get. Adobe was one of the best in terms of type of work, company culture, managing team, perks, and
perks. People: Everyone was very friendly, dedicated, and happy. Intranet/Work Environment: Another luxury I didn't realize I had at the time.
```

Figure 7: Text Summarization of Quora Posts

4 Individual Contribution

As provided in the Project proposal, entire modules are divided and implemented by each contributor whose responsibilities include pre-processing, model training and testing and model enhancements. Detailed module division is as shown below:

Task	Done by	Sub-tasks
Company Review Analysis	Moyank Giri Aishika Nandi	Moyank Giri: Aspect Extraction, Aspect Based Sentiment Analysis (ABSA) Aishika Nandi: Web scrapping, Summarisation models
Company Reputation Analysis	Moyank Giri	Web-scrapping, Feature Engineering, Summarization Models (Pre-trained and Fine tuned)
Exployee Work Life Analysis	Aishika Nandi	Web scraping, Feature Engineering, Trained and Pre-trained Summarization models

5 Conclusion

The above project is structered and implemented as presented in Figure 1. The implementation details and results are provided in the GitHub Repository. Equal contributions are made by both the contributors of the project