IIT BHILAI

MACHINE LEARNING

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

Final Project Proposal

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

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. This shift is evident from the increasing turnover rates within companies, as individuals are now inclined to gain a more comprehensive understanding of their potential workplaces. The challenge lies in bridging the information gap, providing candidates with comprehensive insights into both the advantages and drawbacks of specific companies. This, in turn, enables them to make informed and strategic career decisions that promote long-term stability and job satisfaction.

Our project endeavors to address this pressing 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 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.

Our solution goes beyond mere aggregation by employing sophisticated natural language processing and sentiment analysis techniques. This empowers us to not only summarize a company's working conditions but also to delve into specific aspects that are frequently overlooked. By offering a data-driven, nuanced, and comprehensive view of prospective employers, our project aims to equip job seekers with the tools they need to make informed decisions. This, in turn, will foster a more transparent and accountable job market, benefitting both candidates and employers alike.

Literature Survey

The project proposed in this report was a culmination of multiple references of existing research from top conferences, which are detailed in this section.

The methodology for sentiment analysis described in the paper titled "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach" [1] uses a Deep Learning approach for domain adaptation of sentiment classifiers. The authors aim to address the challenge of training a robust sentiment classifier that can be applied to different domains without the need for domain-specific training data. The approach is based on the idea that Deep Learning algorithms can learn intermediate representations that capture underlying factors of variation in the input data. These intermediate representations can help in disentangling the factors that are shared and meaningful across different domains. The authors propose a transfer loss function that measures the difference between the source domain and the target domain. They also introduce metrics such as the transfer ratio and the in-domain ratio to evaluate the performance of the domain adaptation protocol.

Similarly, another paper titled "A Deep Architecture for Sentiment Analysis of News Articles" [2] presents a deep architecture for performing aspect-level sentiment analysis of news articles. The authors combine various neural network models proposed in different deep learning approaches to address specific challenges commonly encountered in analyzing news articles. The architecture is designed to handle typically long and content-specific news articles, which often lead to overfitting when trained with neural networks. It also effectively processes cases where the subject to be analyzed sentimentally is not the main topic of the article, which is a common issue in aspect-level sentiment analysis. The methodology involves the use of convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and word embedding networks. CNNs are used to capture local features and patterns in the text, while LSTM networks are employed to model the sequential dependencies and long-term dependencies in the article. Word embedding networks are used to represent words as dense vectors, capturing their semantic meanings.

Another paper titled "A review on video summarization techniques" [3] provides a detailed literature review on various methods for Video Summarization which include Deep Learning-based Methods (With the advancements in deep learning, there has been an emergence of methods that utilize deep neural networks for video summarization. These models can learn to extract meaningful features from videos and generate summaries based on those features), Reinforcement Learning methods (Reinforcement learning techniques have also been applied to video summarization. These methods use a reward-based framework where an agent learns to select frames or shots that maximize a reward signal, which is typically based on user preferences or predefined criteria), Multi-Model Approaches (Some video summarization methods incorporate multiple modalities such as audio, text, or metadata in addition to visual information. By considering multiple modalities, these approaches aim to generate more informative and diverse summaries.) and so on

The paper "Combination of Convolutional Neural Network and Gated Recurrent Unit for Aspect-Based Sentiment Analysis" [4] describes a methodology for sentiment analysis which involves a combination of Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU) for Aspect-Based Sentiment Analysis (ABSA). The goal is to extract aspects and analyze sentiment polarities towards those aspects in Chinese product reviews.The paper first preprocesses the text data by tokenizing the words using the Jieba tool and removing stop words. Then, word embeddings are generated using word2vec to represent each word as a vector. The reviews are then padded or cut to a fixed length.The proposed model consists of two main components: CNN and GRU. The CNN is used to extract local features from the reviews, capturing important information at different levels of granularity. The GRU is employed to learn long-term dependencies and sequential relations among the features extracted by the CNN.To perform aspect extraction, the model utilizes global-max-pooling and global-average-pooling to extract global features from the entire sentence. These features are then fed into a fully connected layer for aspect classification.For sentiment classification, the model uses the features extracted by the CNN as input to the GRU. The GRU learns the long-term dependencies and captures the sentiment information related to each aspect. Finally, a fully connected layer is used to classify the sentiment polarity towards each aspect.

Another paper for text summarization titled "Summarizing Online Reviews Using Aspect

Rating Distributions and Language Modeling" [5] focuses on summarizing online reviews by analyzing aspect rating distributions and utilizing language modeling techniques. The authors begin by collecting a dataset of online reviews from various sources. They then preprocess the reviews by removing irrelevant information and extracting the aspect ratings associated with each review. Aspect ratings represent the sentiment or opinion expressed by the reviewer towards specific aspects or features of the product or service being reviewed. Next, the authors analyze the aspect rating distributions to identify the most important aspects mentioned in the reviews. They use statistical techniques to determine the significance of each aspect based on its frequency and the corresponding ratings. This helps in identifying the key aspects that should be included in the summary. To generate the summary, the authors employ language modeling techniques. They use a combination of n-gram models and probabilistic language models to capture the relationships between words and phrases in the reviews. This allows them to generate coherent and concise summaries that accurately represent the opinions expressed in the original reviews.

Datasets

ABSA (Aspect-Based Sentiment Analysis) Dataset: The ABSA dataset [link] encompasses a wide array of reviews, aspects, and sentiments, each meticulously annotated for diverse domains. This dataset is organized with precision, featuring distinct subdirectories for each domain, facilitating focused analysis of aspect-based sentiment across different industries.

FABSA (Feedback ABSA) Dataset: The FABSA dataset [link] offers a diverse and mixed collection of feedback reviews. It includes reviews spanning multiple domains, providing a comprehensive and real-world view of sentiment expression across various contexts. Unlike ABSA, the FABSA dataset does not categorize reviews by domains, making it a versatile resource for analyzing sentiment and aspects in a more holistic manner.

Multi-News Dataset: The Multi-News Dataset [link], developed by Fabbri et al. in 2019, comprises news articles and professionally crafted summaries sourced from newser.com. These summaries are written by human editors and contain references to the original articles they summarize. The dataset is available in English and consists of 56,216 documents in SRC file format.

NewsHead Dataset: The NewSHead Dataset [link], developed by Gu et al. in 2020, encompasses 369,940 English stories, featuring 932,571 distinct URLs. Within this dataset, there are 359,940 stories designated for training, 5,000 for validation, and an additional 5,000 for testing purposes. Each news story comprises a minimum of three and up to five articles.

These datasets play a crucial role in our project by serving as valuable resources for training our models in aspect-based sentiment analysis and text summarization.

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

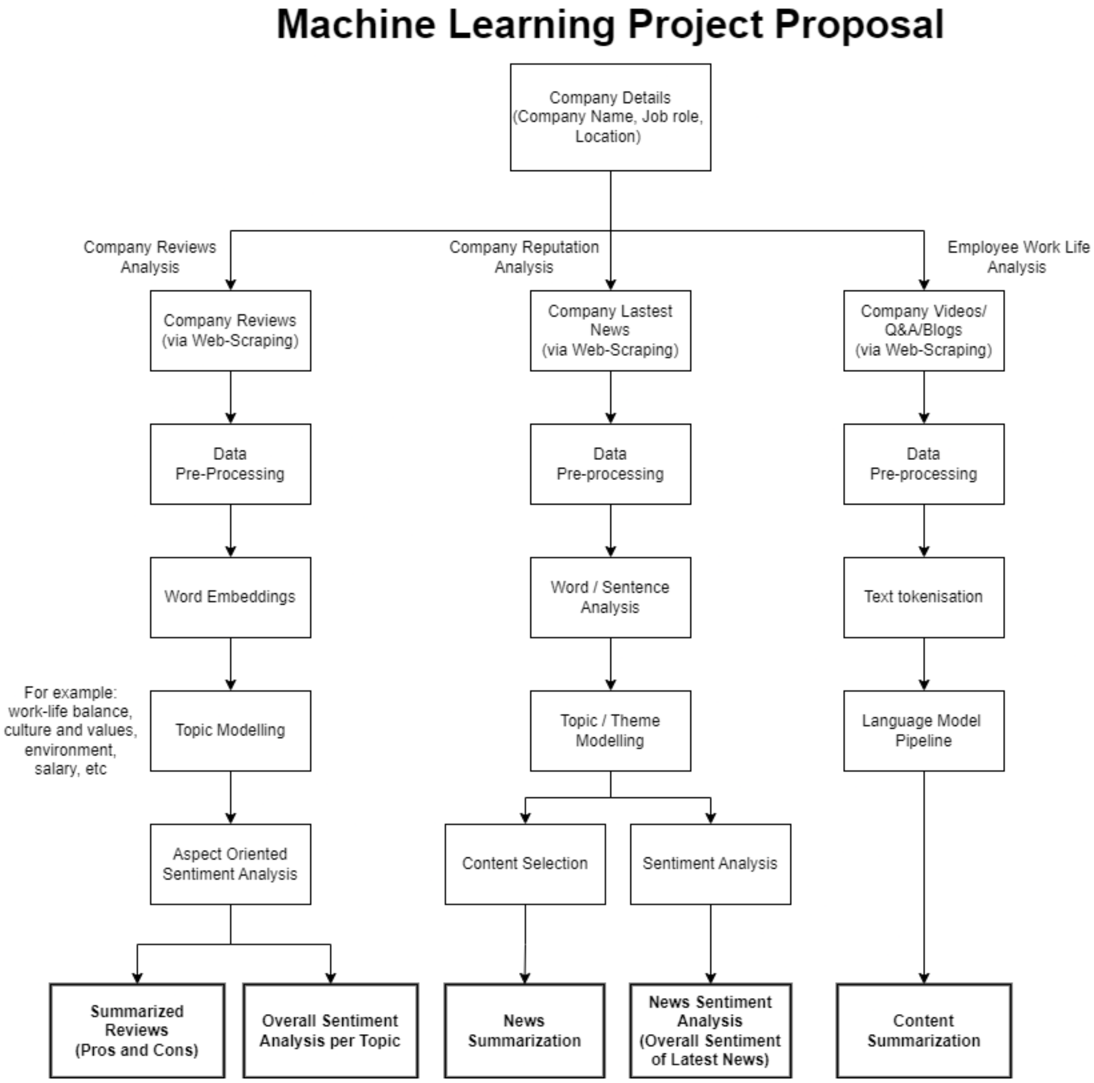
Machine Learning Project Proposal
Company Details (Company Name, Job role, Location)
Company Reviews Analysis
Company Reputation Analysis
Employee Work Life Analysis
Company Reviews (via Web-Scraping)
Company Lastest News (via Web-Scraping)
Company Videos/ Q&A/Blogs (via Web-Scraping)
Data Pre-Processing
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

What is Unique?

The distinctiveness of our project lies in its holistic approach to understanding companies.

- **Reputation Analysis:** Strategic reputation analysis caters to those who seek a comprehensive understanding of a company's current state and want to anticipate future challenges. Users can gauge a company's financial health and whether it's flourishing or facing difficulties, assess job security and the potential risks of consequences such as industry recessions. By aligning with these real-world needs, our project ensures users are well-prepared to navigate their career journey with confidence and foresight.
- **Insights into "Day in the Life":** The unparalleled element of our project is its ability to provide users with an intimate look into what it's like to work at a specific company on a day-to-day basis. While platforms like Glassdoor offer valuable insights, we aim to go a step further by offering detailed summaries of employees' daily experiences by analyzing youtube transcripts, quora posts, and employee blogs, we aim to present a detailed understanding of the work culture, job responsibilities, and daily challenges within a company. This unique feature empowers candidates with a realistic preview of their potential workplace, enabling them to make well-informed career decisions.
- **Diverse data integration:** What sets us apart is our diverse data integration, encompassing news articles for current reputation and summarising personal experiences shared on various blogging sites, knowledge sharing platforms and video transcripts. .

Individual Contributions

The currently planned distribution of work is as follows

Task	Assigned To
Company Review Analysis	Moyank Giri, Aishika Nandi
Company Reputation Analysis	Moyank Giri
Exployee Work Life Analysis	Aishika Nandi

Experimentation Resources

Hardware and Software Resources:
For our experimentation, we will utilize a Lenovo IdeaPad Gaming 3 15IHU6 laptop featuring a powerful NVIDIA GeForce RTX 3050 GPU. This high-performance GPU will enable us to conduct data-intensive tasks efficiently and facilitate the training of machine learning models.

Software Environment:
The experimentation will take place within a Jupyter Notebook environment, a widely used platform for data analysis and machine learning development. We will leverage the versatility of Jupyter Notebook to create, execute, and document our experiments seamlessly. Additionally, we will make use of Python as the primary programming language, alongside various libraries and frameworks for machine learning tasks.

Future advancements

- **Interview preparation tips:** In the future, we plan to expand our offerings by providing in-depth interview preparation tips tailored to the specific requirements of various companies. These resources will encompass a comprehensive breakdown of interview rounds and assessments one has to pass through to crack the company.
- **Skills and certifications common to employees:** As we evolve, we are committed to broadening our horizon by delving into the skills and certifications that are common among employees of different companies. To empower our users further, we plan to implement LinkedIn scraping to extract employee information and analyze skill trends. This will enable us to offer a skill gap analysis, helping users identify areas for skill development.
- **Fake or Real News Detection:** As part of future enhancements, another enhancement that can be applied is to validate the news article before using it in for summarization, which will ensure that no fake information is being given to the user

References

[1] X. Glorot, A. Bordes, and Y. Bengio, “Domain adaptation for large-scale sentiment classification: A deep learning approach,” 06 2011.

[2] D. Nguyen, K. Vo, D. Pham, M. Nguyen, and T. Quan, “A deep architecture for sentiment analysis of news articles,” in *Advanced Computational Methods for Knowledge Engineering*, N.-T. Le, T. van Do, N. T. Nguyen, and H. A. L. Thi, Eds. Cham: Springer International Publishing, 2018, pp. 129–140.

[3] P. Meena, H. Kumar, and S. Kumar Yadav, “A review on video summarization techniques,” *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105667, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197622006571

[4] N. Zhao, H. Gao, X. Wen, and H. Li, “Combination of convolutional neural network and gated recurrent unit for aspect-based sentiment analysis,” *IEEE Access*, vol. 9, pp. 15 561–15 569, 2021.

[5] G. D. Fabbrizio, A. Aker, and R. Gaizauskas, “Summarizing online reviews using aspect rating distributions and language modeling,” *IEEE Intelligent Systems*, vol. 28, no. 03, pp. 28–37, may 2013.