NLP Group Project

APPLICATION DEPLOYMENT – TEAM B(OMBAS)

Timo Bachmann, William Kingwill, Qing Loh, Francisco Mansilla, Alberto de Roni,
Julius von Selchow, Umut Varol

Use case

Justification of the need

Nowadays, job-seeking people are researching an abundance of company information on multiple online platforms such as Glassdoor or Vault. It comes as no surprise that these websites convey different views on a company as an employer, simply depending on the moment of time when they are consulted, or which website they initially consult. With over 28,000 reviews for Microsoft alone on Glassdoor, and reviews listed by date, a reader's opinion is heavily influenced by the first 10 or 20 most recent reviews that s/he will read. Moreover, users are quickly overwhelmed by the sheer amount of data. This challenge is not only faced by job applicants but also from the company side, as an employer. Performing satisfaction surveys, where only current employees are being considered may result in a biased opinion. As per now, companies' sole option to overcome the above issue is to hire a research company to conduct a professional employer brand study. We intend to bridge this gap by providing a tool that efficiently tracks an employer brand image across geographies and various sources, visualizes trends over time, and transmits the real sense of what employees (current, ex, full-time, part-time) truly think and where they see areas of improvement.

Our proposed solution will not only be able to extract textual information of all reviews on one single platform, but also allow aggregation of insights from reviews from a wide variety of sources. Textual information is not simply stored and summarized, but thoroughly analyzed using natural language processing [NLP] techniques. Through a smart combination of polarity and subjectivity scores, we are able to meet such need and extract knowledge from text-based reviews. With regards to the immense collection of company data over several platforms and its processing, there is a clear demand for a well-structured user interface [UI] to facilitate accessibility. In addition, the introduced application can only be successful if the UI allows easy filtering by industry or location, or simply search for a particular company. Everybody who has looked for jobs knows how time-consuming and tedious proper research on future employers can be. What about a unified platform that tackles all the above outlined issues?

Other possible solutions

Another viable approach would be to develop a tool, tailored to each of the mentioned job portals, that would allow a local sentiment analysis of review posts from that respective website. Any platform equipped with such an advanced add-on would have an edge over their competition. Thus, this second use case of providing a licensed use of the tool to online platforms, shifts the original focus of ordinary people as target customer to the website owners. While the general idea remains the same, this approach boasts a higher degree of customizability. Since this product would require tailored solutions for each website, we could focus much more on individual differences and take into consideration specific fields. This is simply not feasible in the chosen, and more general, approach as we focused on scalability of our tool.

Selling proposition / Market justification

As outlined in the first paragraphs, there is a clear need for a tool to store and give access to company profiles as employers. Said application must efficiently condense information from various websites and thereby allowing a translation of textual reviews into some quantifiable score that can easily be summarized. From a more technical point of view, our application applies state of the art NLP techniques to gain maximum insight. The most relevant algorithms in place are sentiment analysis and information retrieval. In the future, the tool will incorporate real-time data streaming. By doing so, we achieve an unrivalled understanding of textual reviews which would otherwise be ignored in a brief company research, or even worse, reading a small subset of reviews might lead to a biased impression. The suggested tool therefore facilitates your personal job research, making it more time-optimal and accurate at the same time.

A second customer group for our application would be for the human resources departments of hiring companies. Human resources will benefit from immediate insights, with significant statistics and visualizations without the need of any coding knowledge. Using the explained tool for their purposes, an HR manager could efficiently gain an idea of the level of employee satisfaction in their own company. Thereby, an interesting addition would be to offer a benchmark comparison, where direct industry competitors' employer profiles are illustrated. This little tweak allows us to scale up this business idea immensely. As a result, the proposed tool can be employed by two different customer groups and embraces two separate use cases almost simultaneously.

Practical implementation

Implementation

The practical implementation in code followed a multi-step methodology. First, the input data had to be retrieved using a customized Python notebook, the Glassdoor scraper (*10_Glassdoor_Scraper_Final.ipynb*). In a second step, all the data collected were merged into one common file and the sentiment analysis was performed. We would like to highlight that our implementation made use of the TextBlob package and not as expected only of the NLTK package. Both have previously been trained on an extensive database. We originally explored the implementation of a classifier of the sentiment of the reviews combining both pros and cons reviews. However, to fulfill our business need of creating a ranking that accurately reflects how the public feels about a company, a mere classification of positive vs. negative is not sufficient. Therefore, we have chosen to use the TextBlob package for our final analysis. As most of our data sources consist of comments and feedback of the public, it can be considered an unsupervised problem where no labels are available. We specifically dealt with this by implementing a way to evaluate the success of our algorithms. More information can be seen from the working code (*1_Overall_Working_Code.ipynb*). Built upon Pattern and NLTK, another advantage of TextBlob lies in its clear documentation and excellent interface. Nevertheless, we still leveraged the original NLTK and Spacy.

Proof of concept – Glassdoor Sentiment Analysis on selected set of companies

To implement a minimum viable product [MVP], or proof of concept [POC], we have selected a small but representative subsample of companies. Thereby, the 2020 PwC list of Global Top 10 technology companies was used as a reference, which included the following:

1. Apple
2. Microsoft
3. Google
4. Facebook
5. Tencent
6. TSMC
7. Samsung Electronics
8. Intel Corp
9. Nvidia Corp
10. Adobe Inc

Per request of an enthusiastic industry expert, the team extended such collection of companies with Datacamp as candidate number eleven.

With this given list of 11 companies, the POC aimed to scrape employer reviews from the Glassdoor website. At this point, the assumption was made that the first 100 pages of reviews would be sufficient to provide a clear picture of the target companies. Based on this assumption, up to 1,000 reviews per company on our list were scraped. Thereby, the following two information chunks from Glassdoor were retrieved:

1. Actual Ratings
2. Review Posts (English only)
    a. Pros
    b. Cons

These would allow us to not only have two reliable sources of employee satisfaction, but also to directly verify the quantification of insights drawn from the sentiment analysis of textual reviews. Specifically, the final score calculated by our application considers the polarity of a review, as the overall polarity of both positive and negative points made, and an additional weight based on subjectivity observed. Since overall ratings are not always available, we can leverage our new Overall Review Score as a proxy for employee satisfaction in case no actual rating had been registered.

Subsequently, we went one step further and crafted a dashboard interface with Streamlit. This allowed us to visualize some key insights with simple Python code, making our minimum viable product a minimum lovable product.

Outlook to full functionality

Taking a more distant view on the discussed application, we envision an all-in-one solution for employer company research that allows users to get a robust overview of their desired company. An intuitive menu will foster easy access to geographical and industry-based drill-down levels in our platform. Leveraging NLP tools such as textual extraction and information retrieval, the company profiles feed from textual employee reviews as an additional source of information. Such reviews will be collected from various sources, including Glassdoor, Twitter, Vault, Indeed, Facebook, LinkedIn, Monster Jobs and Google Reviews. To extend the current scope of already available data, it is important to grasp the dynamics in place with regards to the velocity of data. While a huge portion of reviews are readily available on the mentioned websites, a considerable amount of input data is yet to be published and is uploaded in a continuous manner. That being said, our tool will incorporate a streaming component so as to always ensure that any changes in employee satisfaction are considered. This forms a vital component of the Employer Sentiment Analysis application as opinions may fluctuate quickly in today's world of uncertainty.

We further seek to implement the extraction of number of years worked at a company and use them as weights to compute an average overall score per company. Contributing to the resilience of a key metric, this enhancement strengthens the core of our tool and increases the reliability of the information we provide. As the functionality of the Streamlit tool increases, we would also look to add to the robustness of the dashboard capabilities by adding even more features.

Current field of employer review analysis

As of now, no service is known to us that provides a fully comprehensive review analysis on a single platform. The current state of affairs sees a struggling customer base, i.e. job-seeking people, who have a hard time collecting the information required, but also companies as employers, with widely spread information about their employer brand. Due to limited resources, both from a time and means perspective, potential customers are not able to get a clear picture of the companies they are researching, respectively the employers cannot gather all the information available. Websites such as Glassdoor or Vault already store vast numbers of reviews that would leave one occupied for months for a proper assessment. Moreover, a single source of information may be biased. These take-aways underline the need for a unified platform that conveniently gathers all the information stated in the relevant sources and extracts both numerical but also textual data to calculate reliable employer profiles. In conclusion, this tool could serve as the harmonization of all these websites. Looking into potential development opportunities, companies such as LinkedIn could make use of the tool as a premium service. Both of our target customer groups will perceive high value from such an offer as they could benefit from information and statistics of past employees.