Aspect Based job review Analysis

Pranav A Shenoy,Mohammeed Huvais,Vyshak V

Abstract

- This project aims to help people in choosing the right company for working based on some aspects like learning experience, salary ,work-life balance , management , infrastructure ,workplace.
- The project analyses reviews of a particular company and analyses the polarity for each of the above aspects. The reviews are scraped from indeed.com which contains millions of reviews .Once the scraping is completed, the dataset is preprocessed which includes removing non-english words, punctuations, etc..The reviews are being trained in NaiveBayes and SVM, the features being unigram and bigram.The result is a value (between 0 and 1) corresponding to each aspects, where 0 corresponds to bad and 1 corresponds to good.

Dataset Preparation

- Reviews are scraped company wise from indeed.com using a scraper.These reviews is split into sentences based on full stops, ‘and’ , ‘but’. Each sentence is passed to textblob() to analyse the polarity. Once polarity is analysed, the aspect is extracted and stored in the corresponding file(file name will be that of the polarity). Aspects are chosen based on the count of nouns in the dataset.A

Preprocessing

- Removing Non-english words using enchant dictionary
- Removing stop words like ‘the’, ‘a’ as it does not contribute to sentiment analysis
- Stemming each word so that derivatives of the words will be considered as similar. Eg:opportunity ,opportunities will be considered as similar and result of stemming is opportun

Models For Sentiment Analysis

There are five classes which are neg, sli_neg, neutral, sli_pos, pos.The two models are:

- NaiveBayes model have been trained using unigram and bigram features separately. High accuracy was obtained when using bigrams as features. An accuracy of 67.1% has been obtained for model with unigram as feature and an accuracy of 75.16% has been obtained for model with bigram as feature.
- SVC model have been trained using unigram and bigram features separately. High accuracy was obtained when using unigram as features. The kernel used is ‘linear’. An accuracy of 85.19% has been obtained for model with unigram as feature and an accuracy of 65.25% has been obtained for model with bigram as feature.

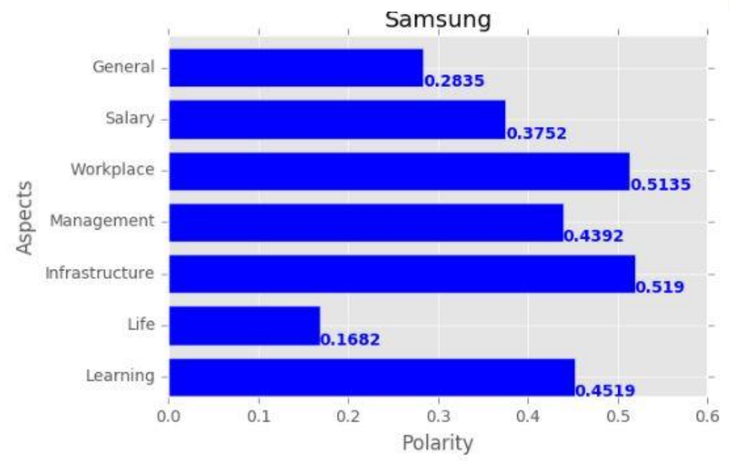
Aspect Classifier

Aspect of a sentence is obtained based on three methodsComparing words of a sentence with a pool of words for each aspects:

- If the words are similar, the sentence belongs to that aspect. In order to improve the accuracy, a word2Vec model has been implemented which increases the size of pool of words .
- NaiveBayes Model: This model uses unigram as feature since aspects could be determined using a single word than n-grams.
- SVC Model : This model uses unigram as feature since aspects could be determined using a single word than n-grams. A linear kernel is used .

Results

- Fresh set of reviews are fetched from indeed.com for analysing. This set of reviews are preprocessed and stored in JSON file . The polarity of each sentence is analysed using the two models and the aspect is analysed using three models.Each polarity is normalised to the range of -1 to 1 (-1 negative and 1 positive) and added to polarity of corresponding aspects. The average value of polarity is taken for each aspect . The total polarity of a company is the average of all the aspect polarity.



Conclusions

In our project we have gone through different methodologies and approaches to do sentimental analysis on a huge corpus of job reviews. We have implemented and analyzed the performance of different approaches involved and have strived to obtain to make the approach as accurate as possible.