Noopur Nishikant Zambare
B20ME051

TASK 4

After keenly observing the implementation part of the system, it was noted that total 600 user-annotated computer science job postings samples were used. Apart from this 10-fold cross validation was used to create training and test sets. Along with it, 540 labeled examples were used to train the extractor. Whereas the additional unsupervised sets varied up to 4000 samples.

The following techniques can be used after the transformation of unstructured data into structured data, would decrease computational expense and improve the prediction results.

1. The most important step to begin any kind of model training is the **Data Preprocessing**. Outliers need to be removed from the data. Empty data cells have to be replaced with mean values and features have to scale uniformly.
   In case when data is scarce, as in this case. Extension of the dataset can be done using pool data from the other sources. Like in the template discussed in paper, if past work experience and qualities would also have been added, then it would have helped in extending the feature set.

2. **Select the components that matters the most or has the highest weightage**
   To select only the most required features various techniques such as PCA and LDA could be used. This filters out the best impacting features and increases computational efficiency at a good extent. But this sometimes leads to elimination of the data. Hence this could work best when the data set is limited or the data samples are less. Another method is to analyze the data and choose the features that are keenly related to the domain and eliminating the rest.

3. **Sending Reduced and relevant features into the Data Mining implementation**
   Data Mining basically comprises of:

   1. DATA AQUITISATION
   2. DATA PREPROCESSING
   3. MACHINE LEARNING ALGORITHM
   4. PATTERN EVALUATION

5. KNOWLEDGE REPRESENTATION

As discussed in the paper the third step of data mining requires a machine learning model to be selected. In this, Decision Tree algorithm is implanted over the data. As per the discussion regarding the cons in the previous tasks, Decision Tree can be updated without actually changing the model the paper had worked upon.
It would definitely help out if decision trees are added with an ensemble of decision Trees. Even, boosted trees are also good alternative.

Boosting includes adding learning algorithm in series to eventually obtain a strong learner. In this initial algorithms are comparatively weaker. In case of gradient boosted decision trees algorithm, the weak learners are decision trees. Every trees goal is to minimize the errors of the tree before it.
Trees in boosting are weak learners but adding many trees in series and each focusing on the errors from previous one make boosting a highly efficient and accurate model.

4. **Changing the implemented algorithm from C4.5 algorithm to C5 algorithm**

Improvement that needs to be done on the basis of algorithm is instead of C4.5, now C5.0 can be used. This would work similar to the previous algorithm but would have few other advantages too.
For example:

- Speed - C5.0 is significantly faster than C4.5
- Memory usage - C5.0 is more memory efficient than C4.5
- Compared to C4.5, C5.0 generates smaller decision tress giving similar results.
- C5.0 has allowed to weight different cases and misclassification types.

5. **RAPIER System**

As discussed, this paper has used simpler version of RAPIER that only applies constraints on the word and part of the speech. However a full-fledged version would perform well.

Above this if few more filters like:

1. **Named Entity Recognition**: NER recognizes the entities such as locations, people, institutions, calendar related terms, etc. from the text.

2. **Sentiment Analysis:** It is the most useful in cases such as reviews, surveys, people's comments and the platform where some feedbacks could be expressed.

3. **Text Summarization**: There are two broad approaches to this, they are extraction and abstraction. Extraction summarizes the text by breaking the chunks of the large data. Abstraction create summary by generating fresh text that conveys the crux of the original text.

4. **Aspect Mining:** It identifies different approaches of an user like it could be positive, negative or neutral or even could be some other possibilities with respect to the document.

5. **Combining several models**

There are high chances of getting more accuracy when results from the one or more models are combined. Considering the example, a final prediction calculated as a weighted average of predictions from various individual models will have comparatively lower variance and improved generalizability compared to the predictions from each individual model. Also, hyperparameter tuning can be done with various models before averaging out.

NEW METHOD

The paper discussed here used Data Mining that in turn uses Machine Learning decision tree algorithms. Neural Network too can be used in such scenarios; the proposed update is based on the implementation of Neural Networks in benefitting the Information Extraction.

The Neural Networks discover the rules from the data itself and eliminate the requirement for the hand crafted rules. Neural networks do not require human intervention as the nested layers within pass the data through hierarchies of various concepts, which actually enables the self-learning through its own errors.

Steps of IR and its broad algorithm

- General input (query or prefix) representation

- General candidate (query or suffix or document) representation
- Estimate the relevance based on input and candidate representation

How Neural Networks can help

- It could help in learning a matching function over the traditional feature based representation of query and document by learning good vector representation.
- Implementation of Embedding
  Usually embedding captures the semantics of the input by placing semantically similar inputs close together in the embedding space. An embedding has advantage that it can be learnt and reused across models.