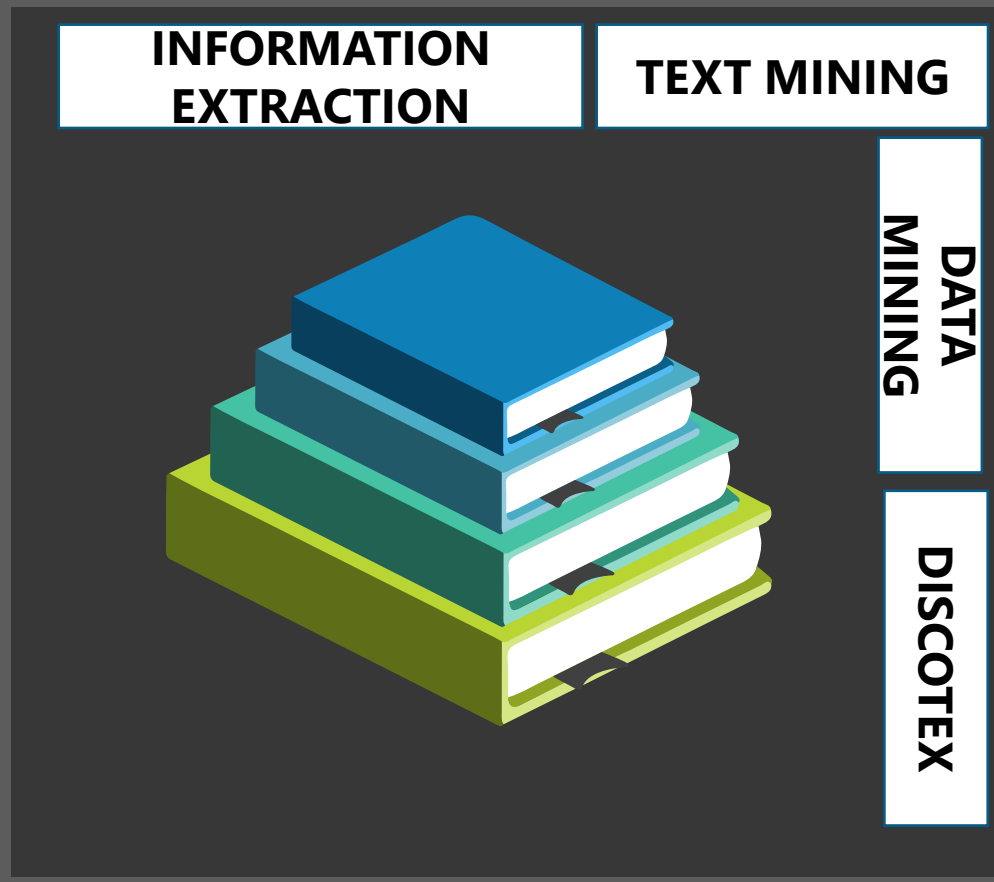
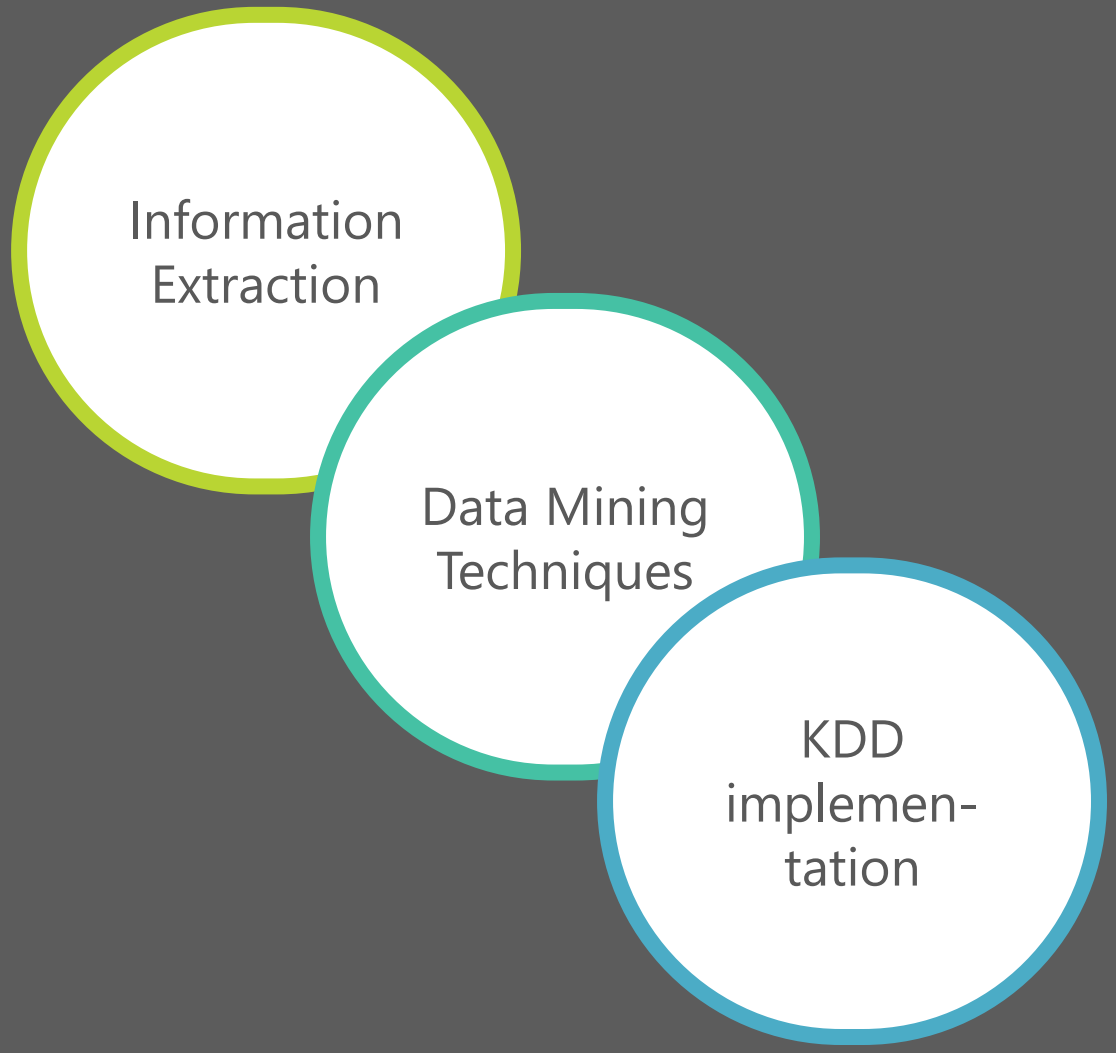
Name- **Noopur Nishikant Zambare**
Roll no.-**B20ME051**
Branch- **Mechanical Engineering**

Topic-'**A Mutually Beneficial Integration of Data Mining and Information Extraction**'
Research paper by: **Un Yong Nahm and Raymond J. Mooney**



What does this paper emphasize on ?



Basically this paper has aimed on

- Increasing the number of rules for prediction by implementing RAPIER system and then DiscoTex over it.
- Relational patterns are created among the features by treating them as a slot.
- Hence it can be said that slot fillers have helped in improving the predictions.

What Machine Learning algorithms this paper has used ?

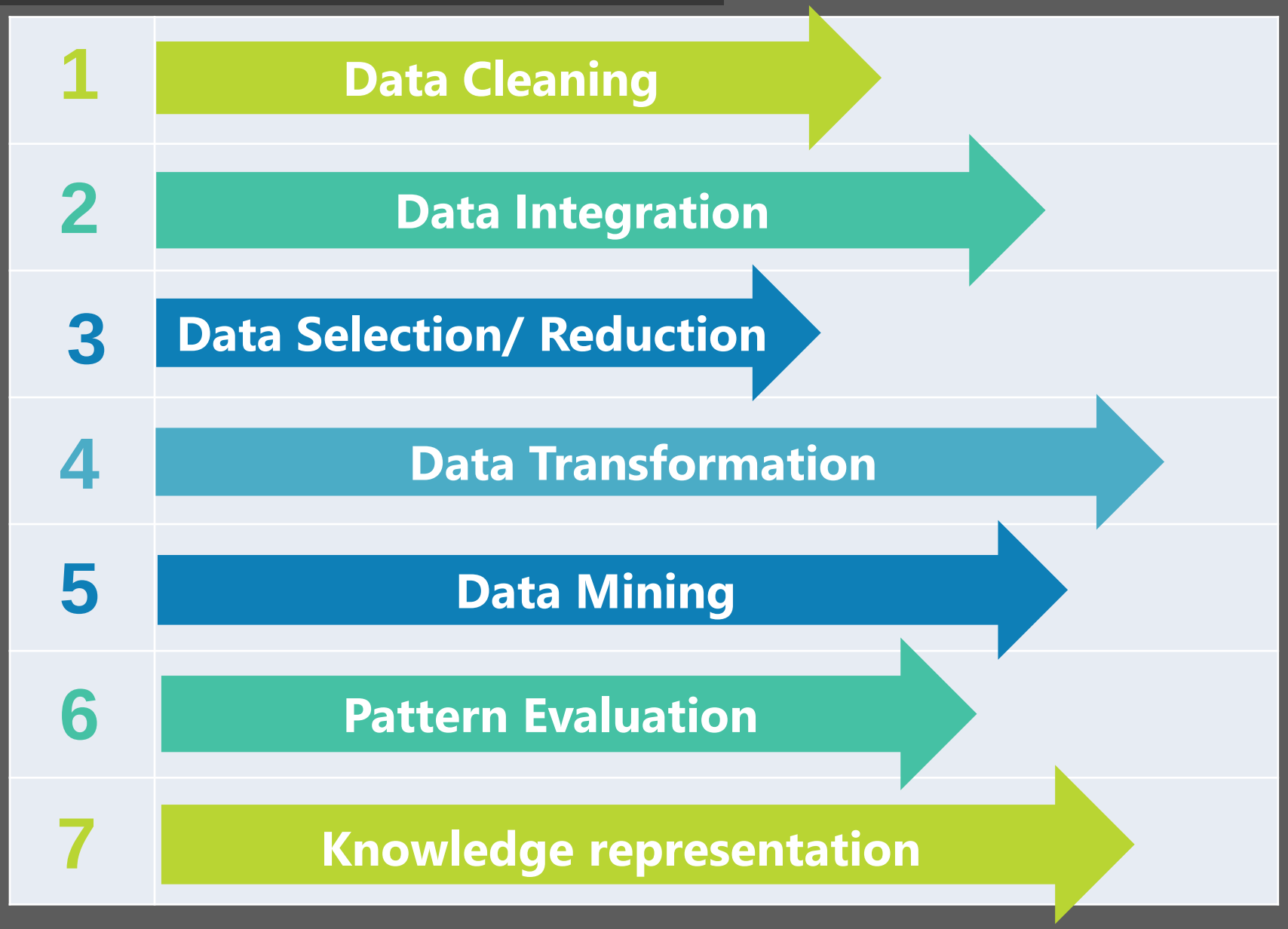
- In this paper, DISCOTEX uses C4.5 rules to induce rules from the resulting binary data by learning decision trees and translating them into rules.
- Discovered knowledge describes the relationships between slot values is written in the form of rules.
- It is a Conditional Control Model as Desicion trees are used.

- Basically the ML model is used while applying data mining algorithm. It works on the model that is in turn creating judgment by forming Decision Trees.
- Moreover some mathematical concepts like recall, precision and F-measure are also discussed in the paper for evaluation purpose.

KDD-KNOWLEDGE DISCOVERY IN DATA

KDD is
an repetitive process where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.

- It is knowledge extraction or information extraction from the data.



01 DATA AQUITISATION

02 DATA PREPROCESSING

03 MACHINE LEARNING ALGORITHM

04 PATTERN EVALUATION

What is Data Mining?

Apart from KDD what Data Mining exclusively consists is the Machine Learning Algorithm. The preprocessing and the evaluation steps resemble that of KDD with including the model such Decision Trees context to this paper.

Information Extraction

Understanding relevant and useful part of the text

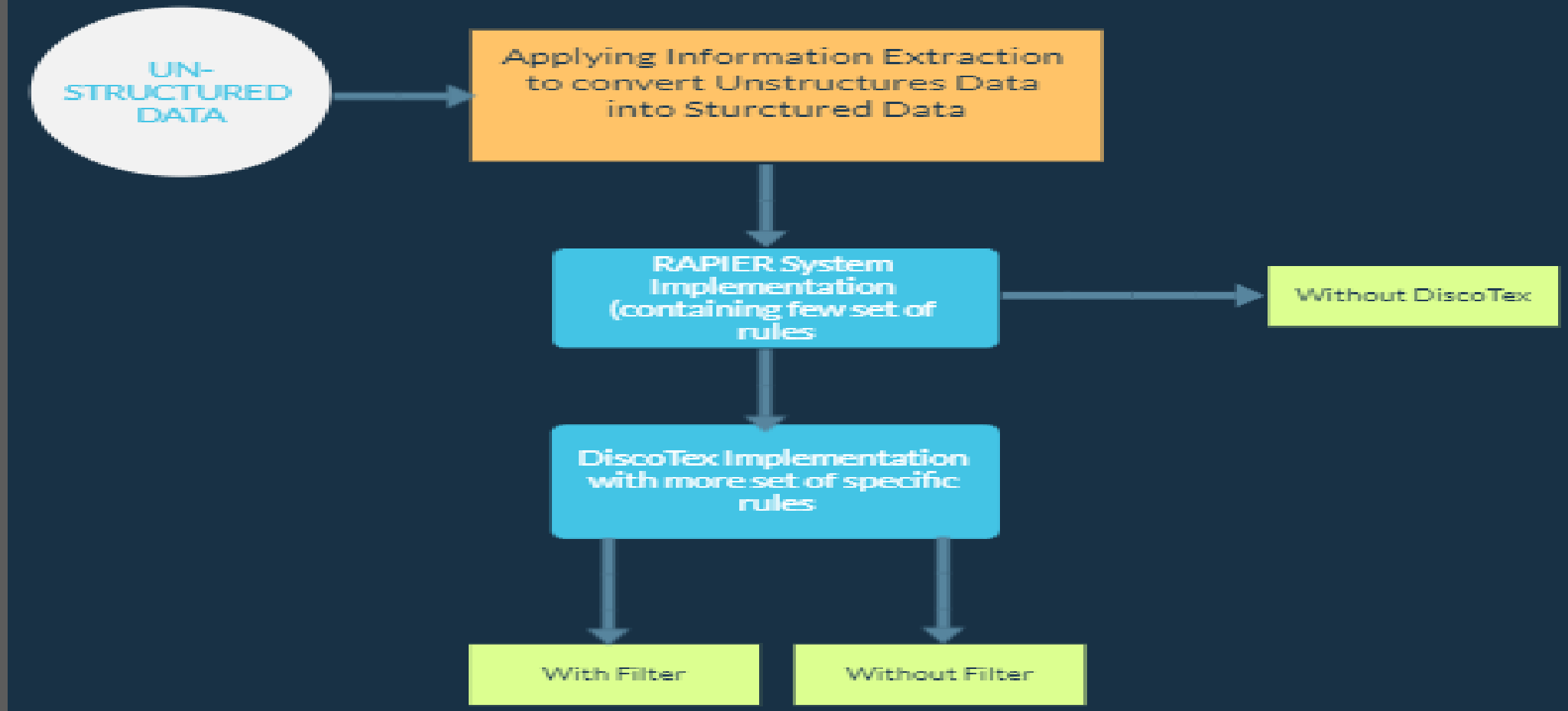
Gathering information from huge text data

Creating a structured information of the relevant data

Goal

- Information organization for ease of understanding
- Arranging information in a semantically precise form that would allow inference to be made by computer algorithms.

HOW DOES DISCOTEX WORKS



Different Approach

Data Preprocessing

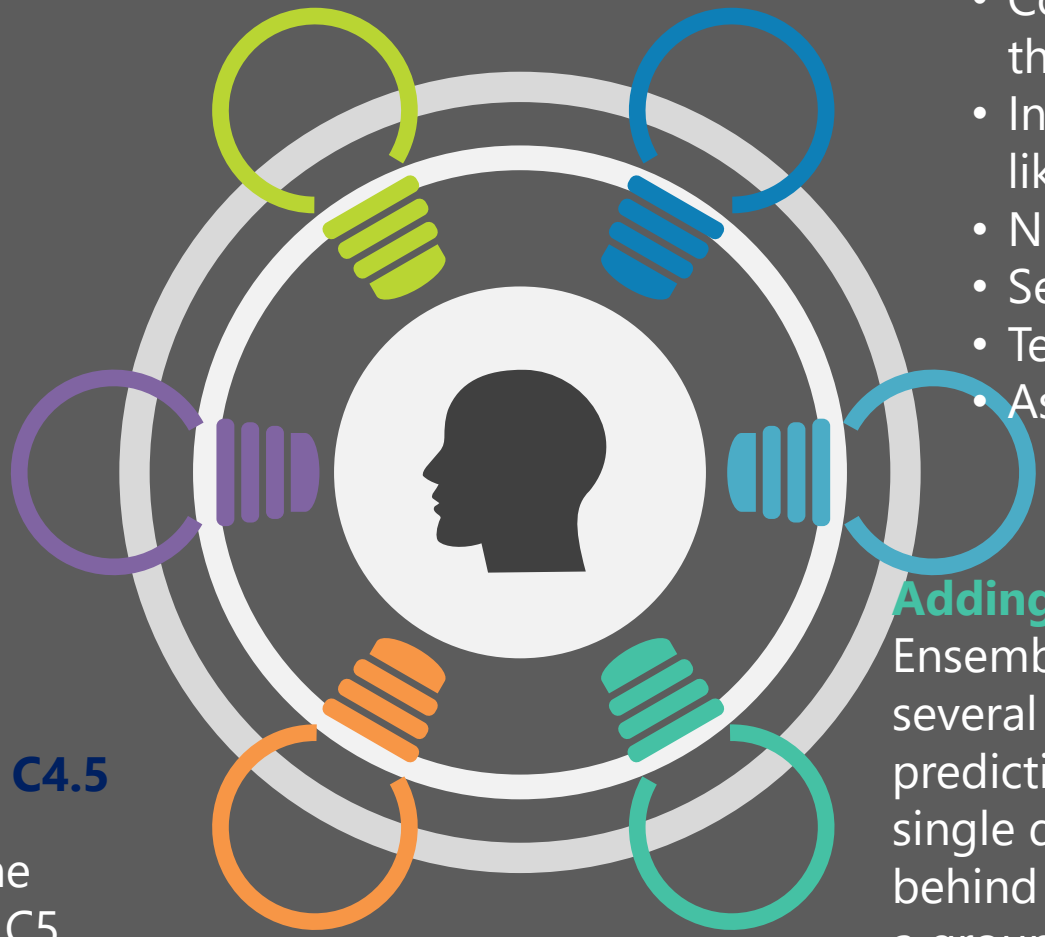
- Extension of data set using pool
- Removing outliers
- Replacing missing data.

Applying PCA or LDA

When the data set is large or only relevant features are required, then these techniques work well.

Using C5 algorithm instead of C4.5

During the year 2000 C4.5 was the updated one but within 21 years C5 algorithm has solved the issues of previous one and performs better.



Modifications in RAPIER System

- Considering a full-fledged system that a simpler version.
- Inducing other filter components like :
- Named Entity Recognition
- Sentiment Analysis
- Text Summarization
- Aspect Mining

Adding Ensemble Decision Trees

Ensemble methods, which combines several decision trees to produce better predictive performance than utilizing a single decision tree. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner.

CONCLUSION

For the final evaluation, over 600 job posting templates were used along with 4000 unsupervised labels to discover hidden relations. Performance was judged by calculating recall and F-measure value.

APPLICATION

- **Analyzing reviews**
- **Email detection**
- **Business Analysis**
- **Bio logical data analysis**

3 models were created:

1. RAPIER System
2. RAPIER System with DISCOTEX but without filter
3. RAPIER System with DISCOTEX with filter

F-measure for DISCOTEX gave almost similar values irrespective of the additional filter. But there is significant variation between non DISCOTEX and the Model that paper described.