**Assigned Tasks:**

Task1: Done (future work preparing a set of words which are significant and provide better resembles the characteristics of the deal)

Task2: Done(clustering or grouping based on similarity once again we can have better definition for this problem if we manually study through good deal and bad deals data)

Task3: Partially done.

Areas of betterment in future work.

1. In Task 2 and Task 3 further development can be done using advanced algorithm for feature extraction and classification, further details and status of the project is provided below.
2. Deep learning algorithms with better features would provide optimal solution.

**Feature extraction method used in this project**

Method1: Using vowel frequency as vector, so you get a five feature vector for each data point. (Performed but results were not as expected, python.)

Method2: Statistical features derived from converted text into ascii (Mean, variance, median, and entropy). (Performed using Decision trees)

Method3: NLP and text blobbing (Currently working)

Method4: Co-occurrence method (Performed using NN and SVM, AUC of 0.98 and error rate of 8% in training)

**Future works on feature extraction**

Method5: Feature selection methods (wrapper method and classification guided feature extraction)

Method6: cluster based method for selecting significant features (key words)

**Data organization:**

1. Training data is divided into 70% for training and rest of the 30% for validating
2. K-fold cross validation with montecarlo repetition.(Due to lower number of training samples only monte carlo repetition was implemented )
3. System is blindly tested on trained system. (Implemented using Decision trees, NN and SVM’s)

Classification:

1. LDA(Did not perform well), Naïve Bayes(marginally worked),

### Neural Network (Implemented using Bayesian back-propagation, scaled conjugate and Levenberg-Marquardt back-propagation)

1. SVM(Performed using Linear kernel)

Classification modes:

1. Individual classification performance(Performed )
2. Mixed classifier(Performed)
3. Ada-boosting(Future work)

Fusion: (future work)

LDA & NN (Simple sum rule, min rule, and max rule)

SVM (Voting based on various ensembles)

Metrics:

1. Area under the curve (AUC)(Performed in this study)
2. Specificity and sensitivity(Will be reported for other types of kernels in future study)

Q& A at the time of solving the problem

1. How can we build a system that extracts structured data from unstructured text?
2. What are some of the robust methodologies for identifying the entities and relationships described in the text?
3. Which corpora are appropriate for this work, and how do we sue them to train and evaluate other models?

Possible solution

1. Data stored as tuples for example (entity, relation, entity)
2. It is really hard for the data to be structured, so one best answer would be using relationship between entities.
3. Information extraction\*\*\* Information Extraction has many applications, including business intelligence, resume harvesting, media analysis, sentiment detection, patent search, and email scanning\*\*\* in our case deal detection.
4. Chunking: segments and labels multi tokens (feature extraction based on weights (specify characterization)), ex n-gram approach.
5. Text blobbing(will be solved in future versions)

* Noun phrase extraction
* Part-of-speech tagging
* Sentiment analysis
* Classification (Linear and nonlinear classification)
* Tokenization (splitting text into words and sentences)
* Word and phrase frequencies
* Parsing
* *n*-grams
* Word inflection (pluralization and singularization) and lemmatization
* JSON serialization
* WordNet integration

Observations:

1. Items not considered

* Numbers
* . , / \ - ‘’ ! # $ + & : | .com
* Stop words
* Repeating words