Financial Annual Report text classification

Presenter : Xuan Liu
Advisor : Professor Yingshu Li

2017/12/01

Outline

- Motivation
- Dataset & Class Definition
- Feature Selection
- Algorithm-level Comparison
- Assessment Metric
- Result & Analysis

Motivation

- Customer-centric structure are exhibit superior financial performance compared with firms that are internally structured (mostly product-centric)
- Machine Learning tools may help automatically determine company's management alignment and marketing strategy.
- Hard to be found on company's website and newsletter but can be retrieved from firm's annual report 10-k filing.

Dataset & Class definition

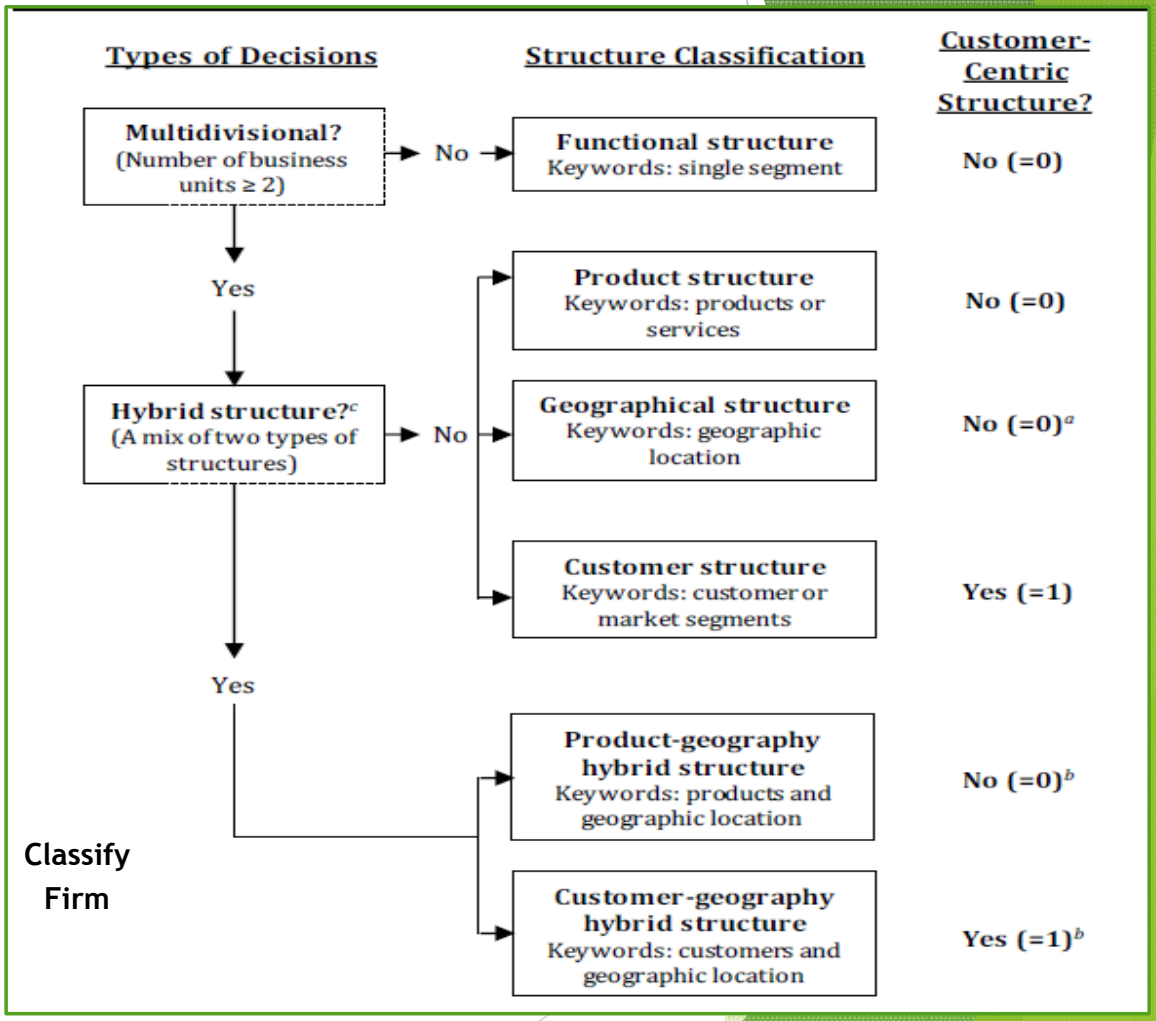
- Attributes and Business Insight
  - Highly structured official documents with large hidden information.
  - Plain Written English
  - Textual data combined with financial numerical data

- Target Portion
  - Management's Discussion and Analysis (MDA)
    - MDA is an important document for analysts and investors who want to review the company's financial fundamentals and management performance.

Binary Classification

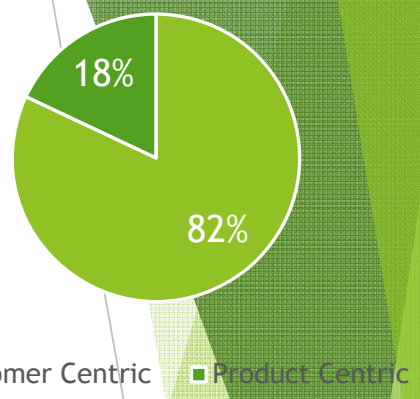
- Customer Centric vs Product (Service Centric)
  - Customer-Centric alignment on corporate level can be detected by checking their segment note included in 10-k reports.
  - Our approach is to generalize the classification using MDA context.

Segment Section — Supervised Learning — MDA Section

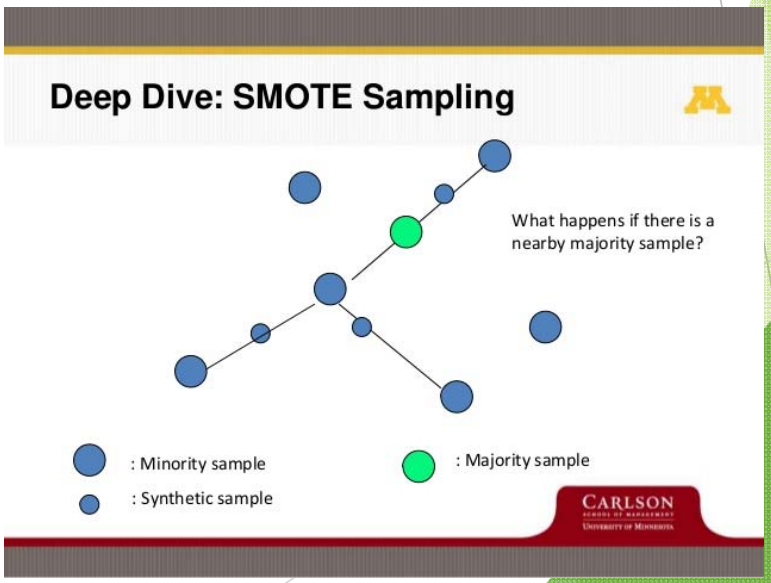


Dataset Class Distribution

- Use annual report required section “Segment Information” which reveals the internal structural corporate design to label the whole data set.
- Year 2016, whole annual report pool contains 7500 firms’ with complete text version report.



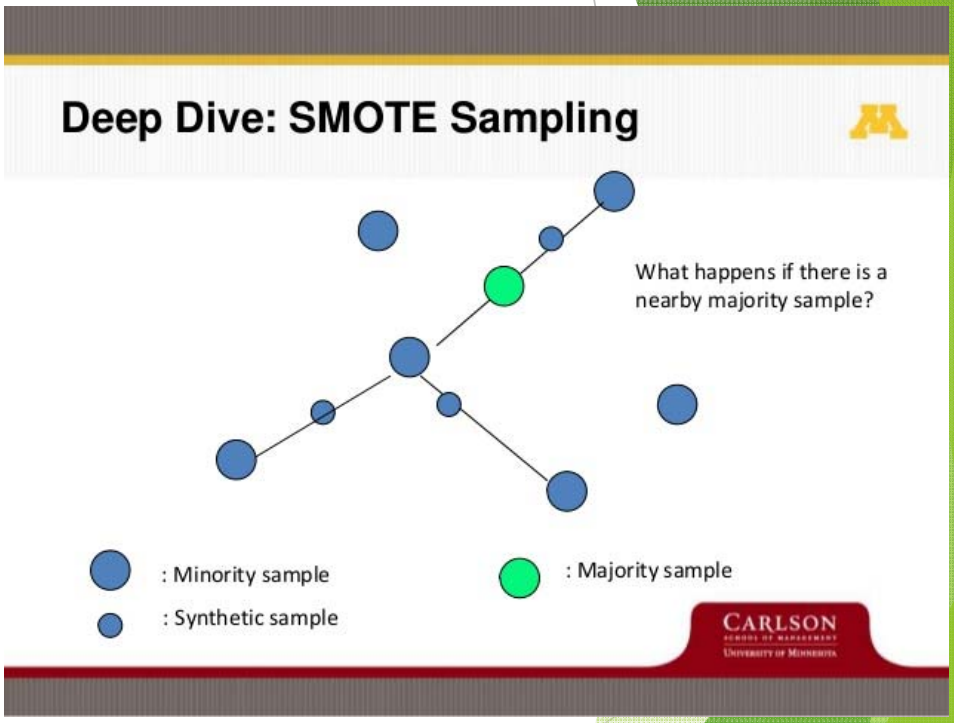
- Resampling (SMOTE) vs Non-Resampling



Resampling (SMOTE) vs Non-respelling

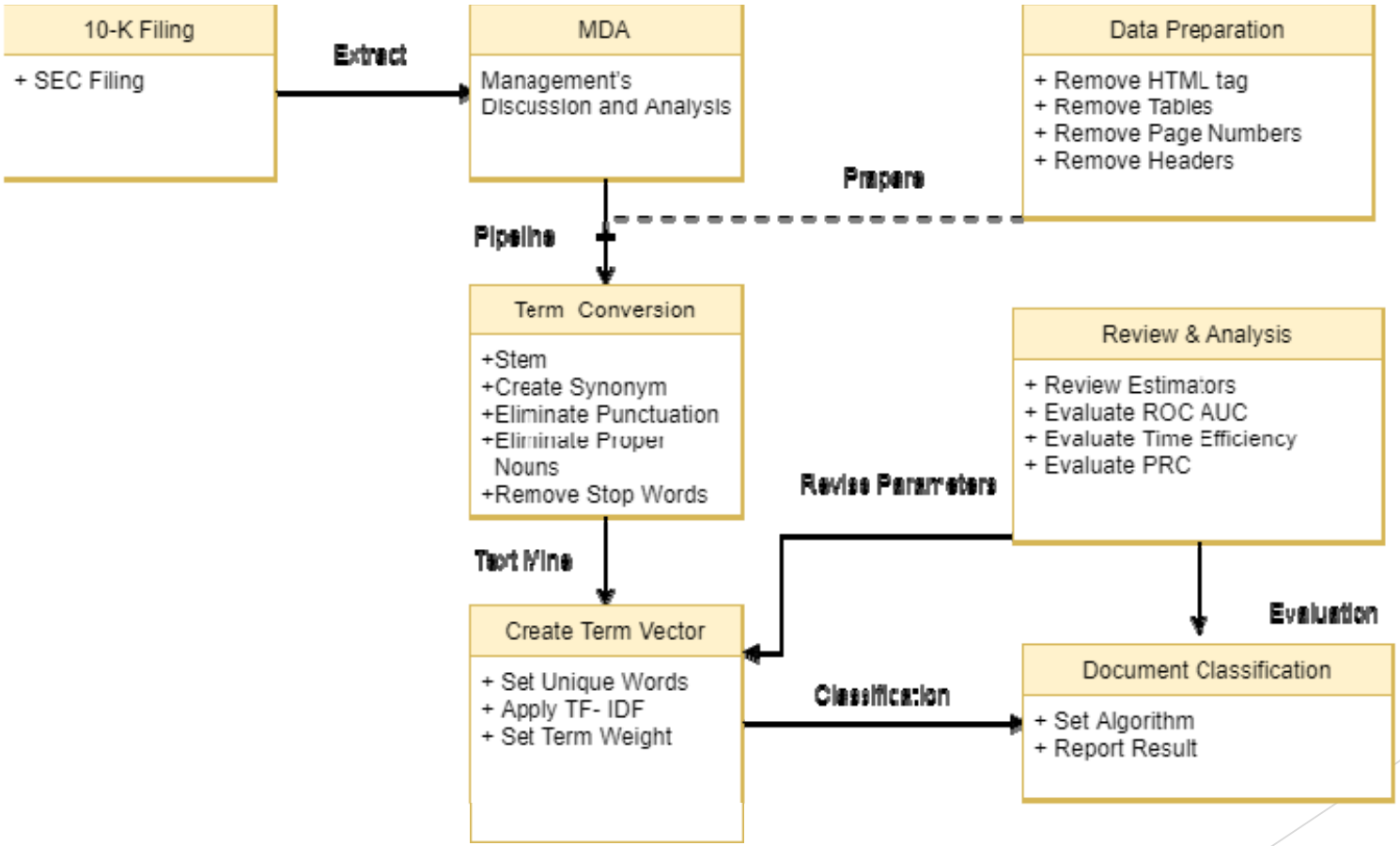
SMOTE –synthetic Minority Oversampling Technique

-Combines informed oversampling of minority class with Random undersampling of majority class



AUC	Non-resampling	SMOTE
LR	0.6928	0.68397
Random Forest	0.6717	0.671258

Data Flow

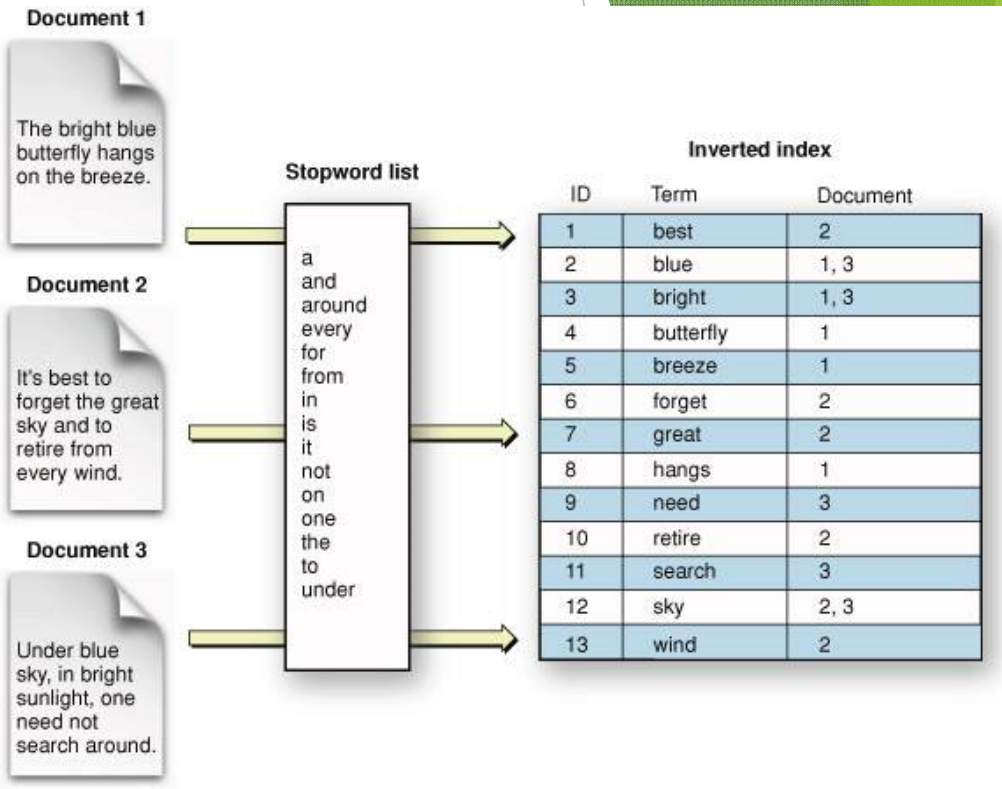

10-K Filing
+ SEC Filing
Extract
MDA
Management's Discussion and Analysis
Data Preparation
+ Remove HTML tag
+ Remove Tables
+ Remove Page Numbers
+ Remove Headers
Prepare
Pipeline
Term Conversion
+Stem
+Create Synonym
+Eliminate Punctuation
+Eliminate Proper Nouns
+Remove Stop Words
Review & Analysis
+ Review Estimators
+ Evaluate ROC AUC
+ Evaluate Time Efficiency
+ Evaluate PRC
Revise Parameters
Text Mine
Evaluation
Create Term Vector
+ Set Unique Words
+ Apply TF- IDF
+ Set Term Weight
Classification
Document Classification
+ Set Algorithm
+ Report Result

Feature Selection

- BOW – Bag of Words
  - Model - a way of representing text data when modeling text with machine learning algorithms.

- **Pros**
  - Easy to computer
  - Has basic metric to extract most descriptive terms in the document
  - Easily compute the similarities between two documents

- **Cons**
  - Based on BoW, it does not capture position in text, semantics, co-occurrences
  - TD-IDF is only useful as a lexical level feature
  - Can't capture semantics



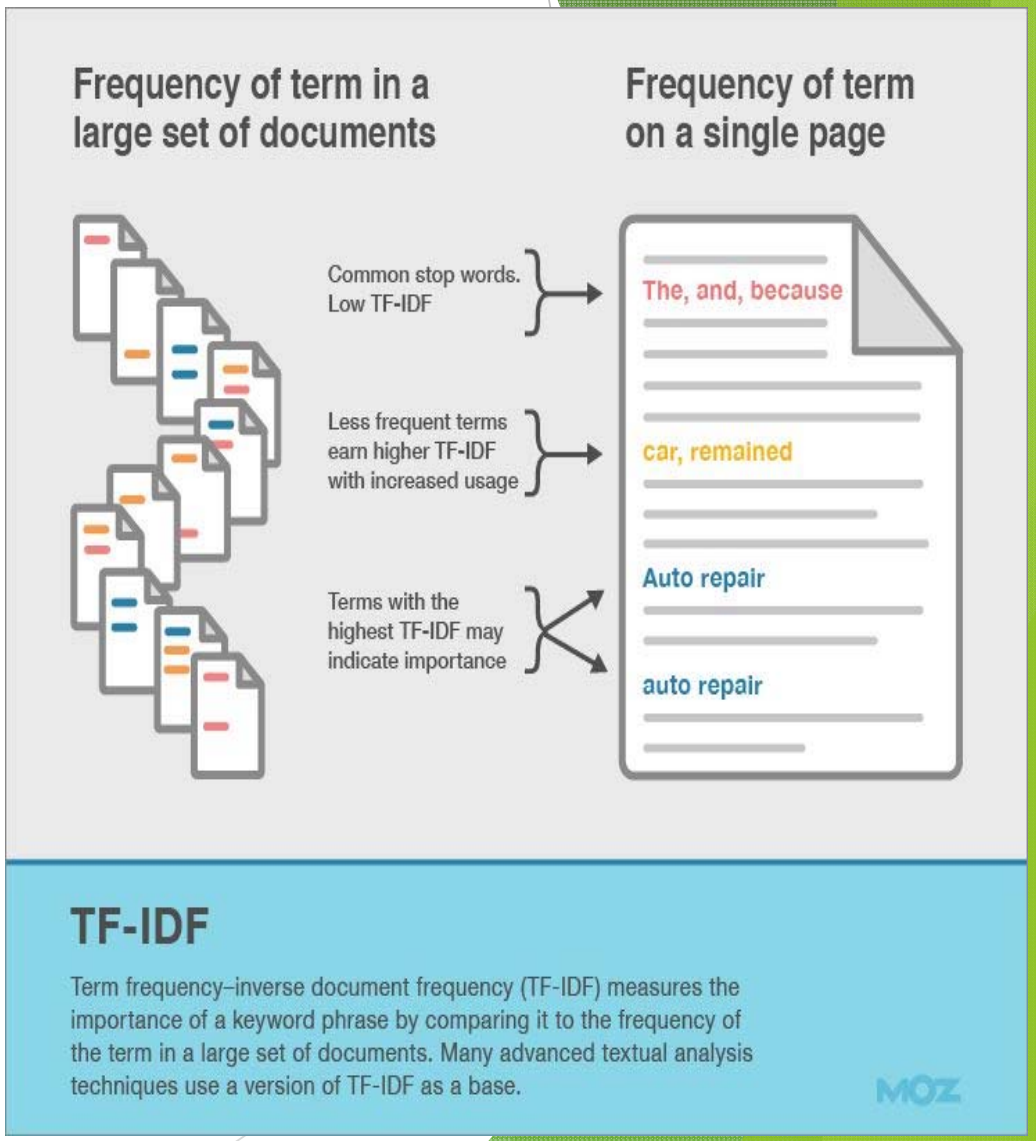
TD-IDF

$$TF - IDF = TF_t \times \log \frac{N}{DF_t}$$

TF_t – **Term Frequency** of the term t. (How many times does the term occur in the document?)

N – Total Number of Documents in the Corpus.

DF_t – **Document Frequency** of t. (How many documents have the term t?)



N-gram

- N-gram
  - Sequence of tokens of length N
  - Can be words, combination of words/terms.
  - Unigram (1-item), Bigram(2-items), Trigram(3-items)

Algorithm-level Comparison

- SVM
- ~~Neural Network~~ (Too many layers to be engineer and time consuming)
- ~~Naïve Bayes (Simple)~~
- Random Forest
- Logistic Regression

- TD-IDF + SVM
- TD-IDF + Random Forest
- TD- IDF + Logistic Regression

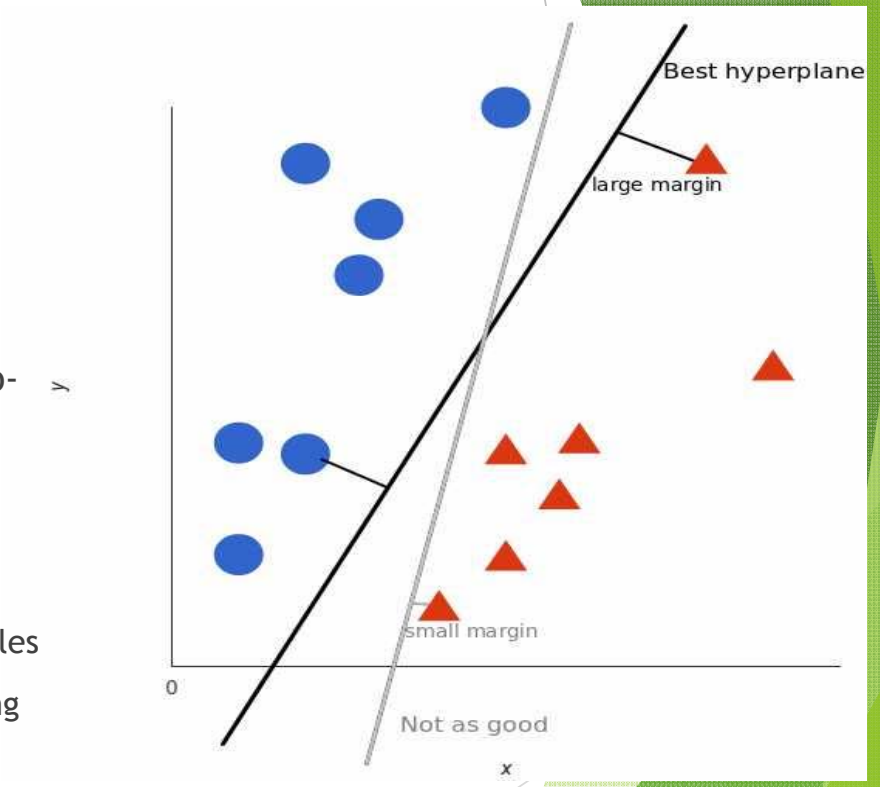
SVM

Support Vector Machines advocates lots of machines as many as all hyperplane that performs

- **Pros**
  - Many public available SVM packages
  - Kernel-based framework is powerful, flexible
  - SVMs works very well in practice, even with very small training sample sizes
- **Cons**
  - No “direct” multi-class SVM, must combines two-class SVMs
  - Computation, Memory

1) During training time, must compute matrix of kernel values for every pairs of examples

2) Learning can take a long time for large-scale problems

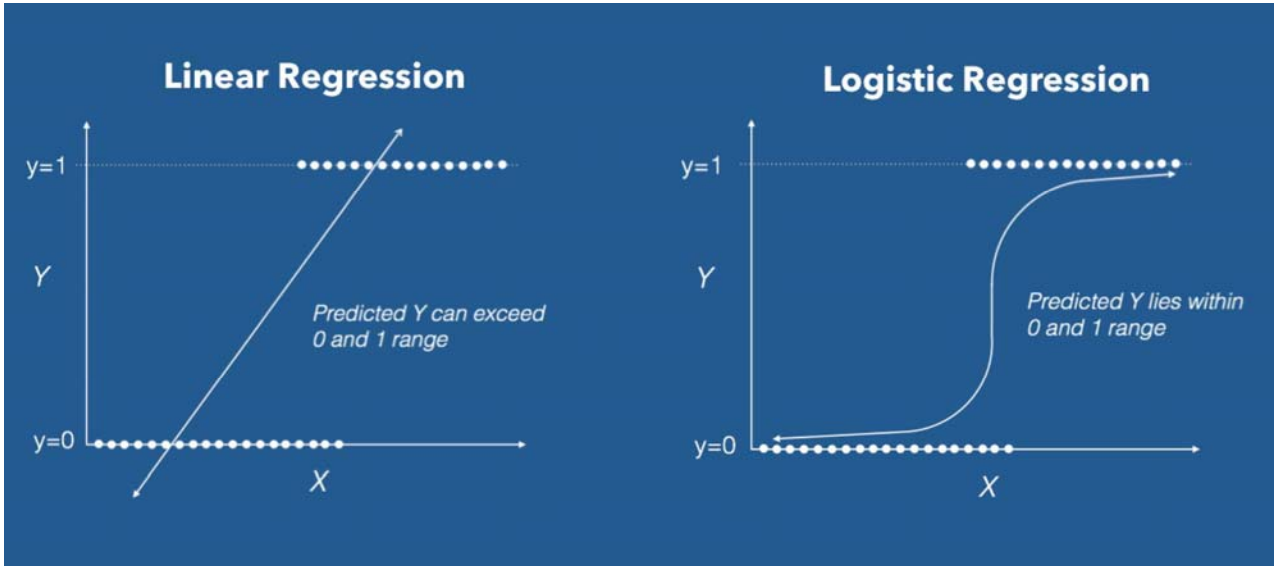


$$\arg\max_{boundary} margin(boundary)$$

Subject to distances of all correctly separated elements belongs to either side $\geq$ margin

Logistic Regression

Logistic regression is a linear algorithm (with a non-linear transform on output). It does assume a linear relationship between the input variables with the output.

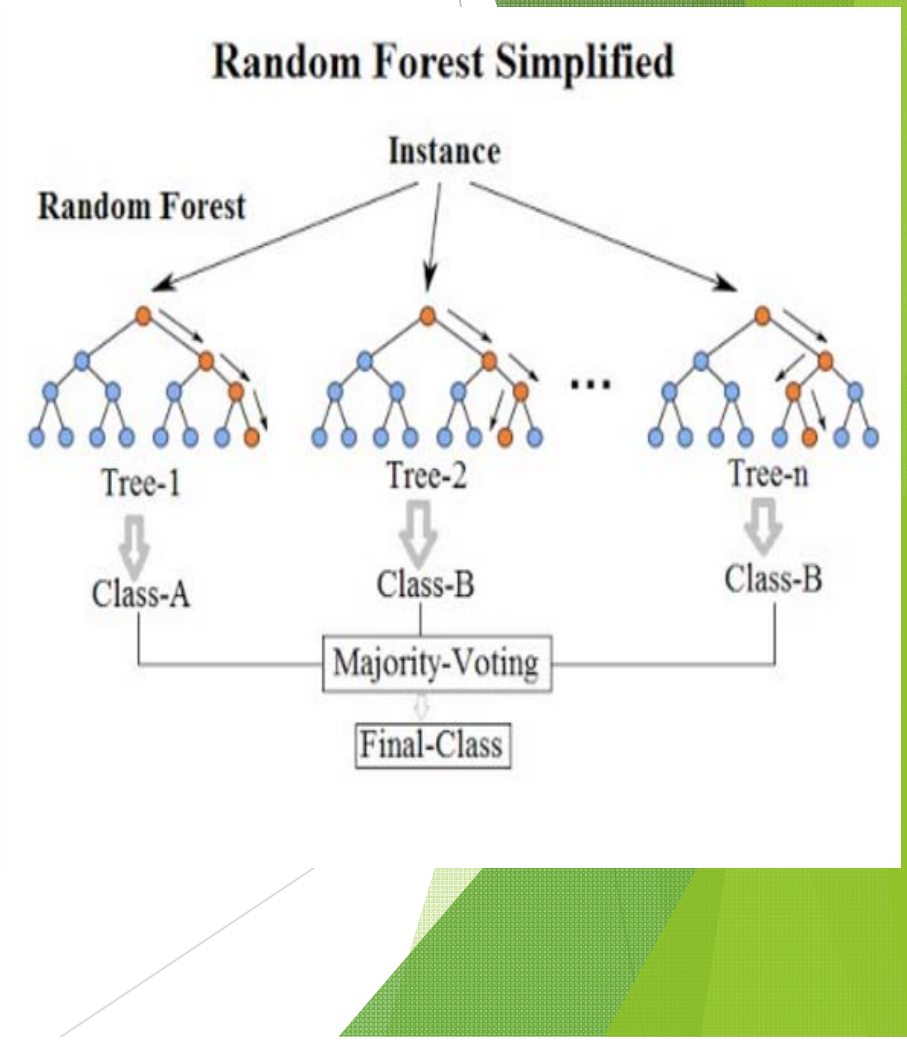


- **Pros**
  - Easy to interpret – the idea of regression is familiar and intuitive
- **Cons**
  - Require Certain statistical assumption to hold true in data
  - Generally low predictive accuracy ?
  - Like linear regression, the model can overfit if you have multiple highly-correlated inputs.

Random Forest (Ensemble)

In a forest with T trees we have **t ϵ {1,…, T }.** All trees are trained independently (and possibly in parallel). During testing, each test point v is simultaneously pushed through all trees (starting at the root)

- **Pros**
  - Generalization through random samples/ features
  - Very fast classification
  - Inherently multi-classes
  - Simple Training
- **Cons**
  - Inconsistency
  - Difficulty for adaption

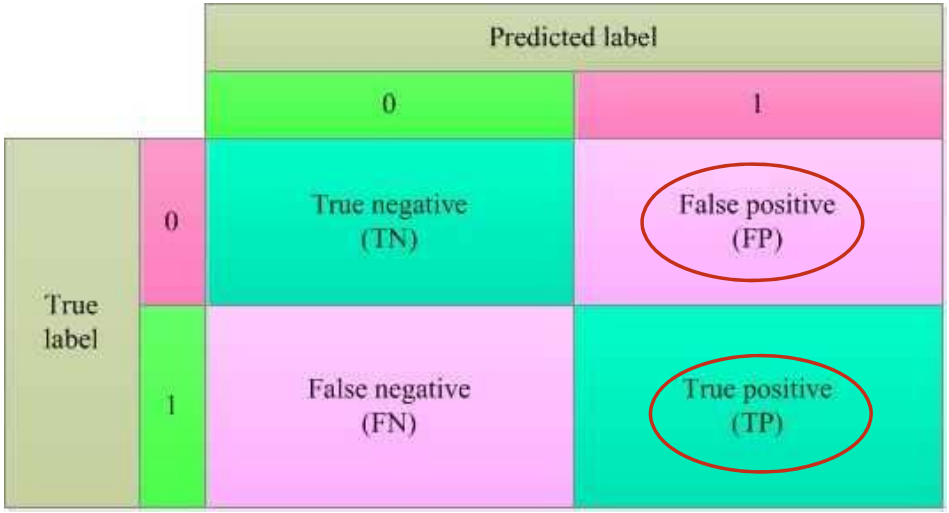


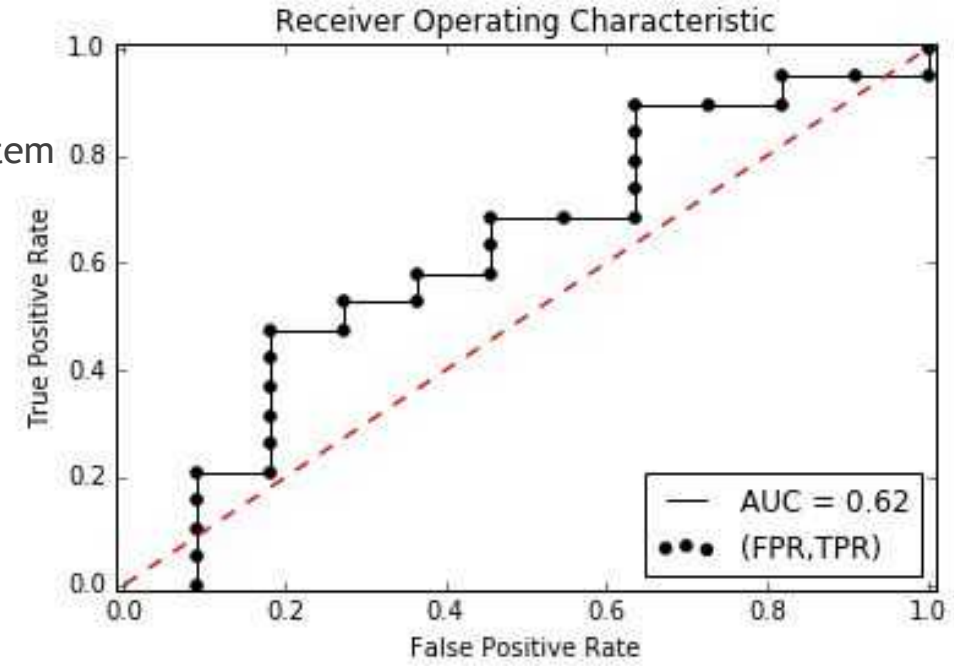
Assessment Metric

- ROC AUC
- PRC
- Time Efficiency

ROC AUC

- ROC – receiver operating characteristic curve,
  - illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
- AUC - Area Under Curve
  - Range [0.5, 1)
  - Demonstrate if a sample is true label, the possibility of classifier determines it's a true should be larger than classifier determines it's a false.



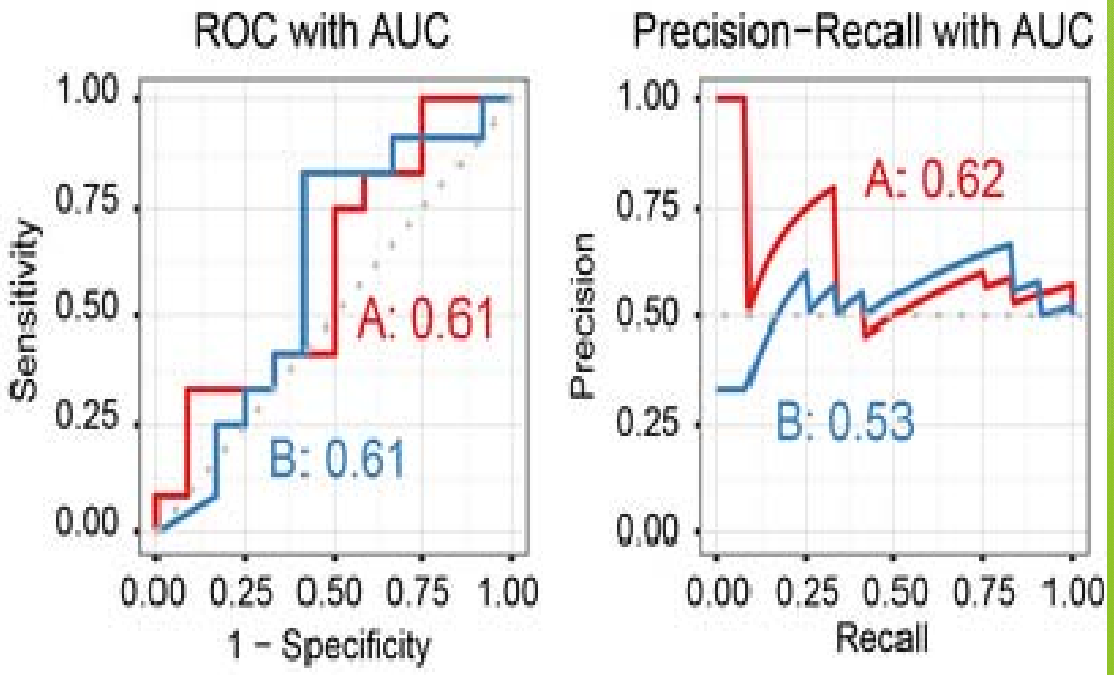


Recall = TPR = TP/(TP+FN)

False Positive Rate = FP/(FP+TN)

PRC

- PRC – Precision-Recall curve
  - **Precision = TP/(TP+FP)**
  - **Recall = TP/(TP+FN)**
- Advantage
  - More sensitive than ROC AUC
    - If we need high recall to detect each positive event.
    - If dataset is not balanced, negative instance is much more than positive instances.



Time Efficiency

time\ (sec)	5000	10000	15000
LR	5.8280	7.196	5.414
RF	49.623	38.508	33.319
SVM	30min+	30min+	30min+

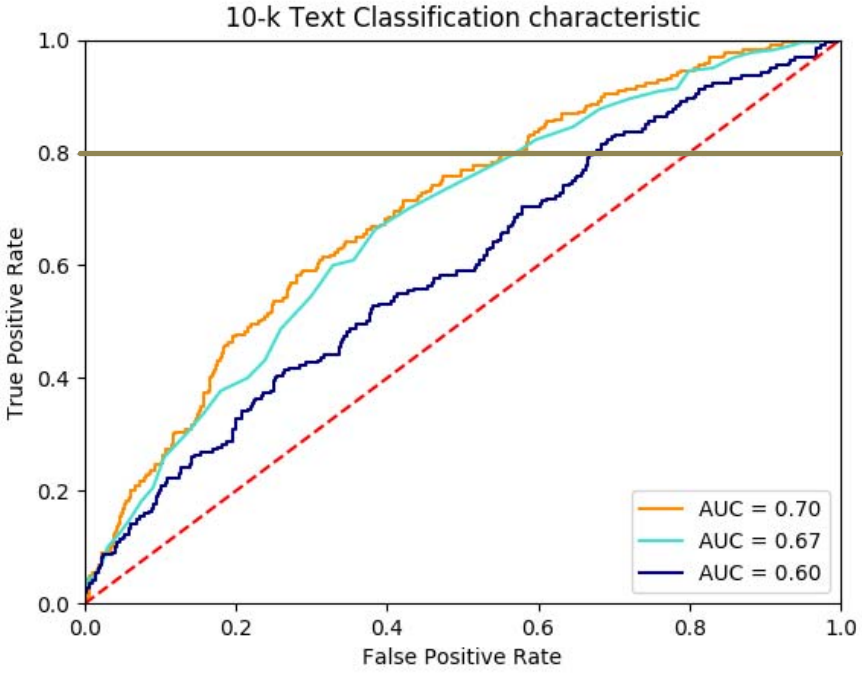
- Logistic Regression Classifier tends to train dataset faster than other algorithms.
- SVM classifier left a way behind other classifiers.
- Random Forest relatively outperforms when number of feature is scaled up.

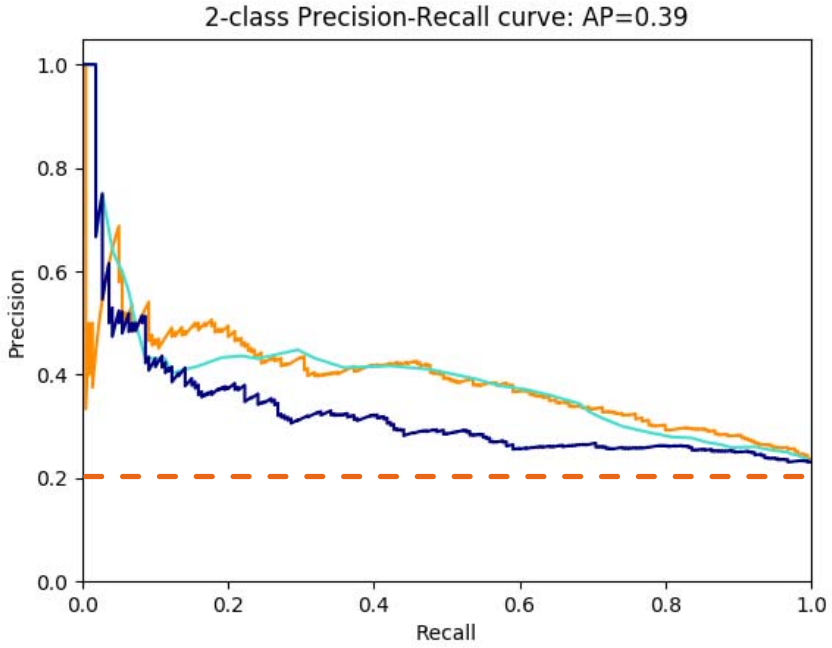
Result & Analysis

- Tables shows 5-scenario on ROC AUC performance

Max-feature	5000(small)	5000(large)	10000(small)	10000(large)	15000(large)
Logistic Regression	0.69327	0.6897	0.6928	0.6957	0.6710
Random Forest	0.6661	0.67029	0.6717	0.6716	0.6788
SVM	0.55521	0.5013	0.58259	0.5955	N/A

ROC VS PRC





Note: No resampling, baseline for PRC = 0.2 is based on positive/negative ration 1:4

Conclusion

- Linear model (Logistic regression) for **sparse high-dimensional data** such as text as bag-of-words.
- Imbalanced dataset use resampling approach outperform compared with not using it.

Reference

- The Relationship Between Precision-Recall and ROC Curves, http://pages.cs.wisc.edu/~jdavis/davisgoadrichcamera2.pdf
- Differences between Receiver Operating Characteristic AUC (ROC AUC) and Precision Recall AUC (PR AUC) http://www.chioka.in/differences-between-roc-auc-and-pr-auc/
- Imbalanced Data Sentiment Analysis in Short Arabic Text
- Effect of Customer-Centric Structure on Long-Term Financial Performance, Lee et al_MKS 2015_org structure
- Exploring the Forecasting Potential of Company Annual Reports

Thank you