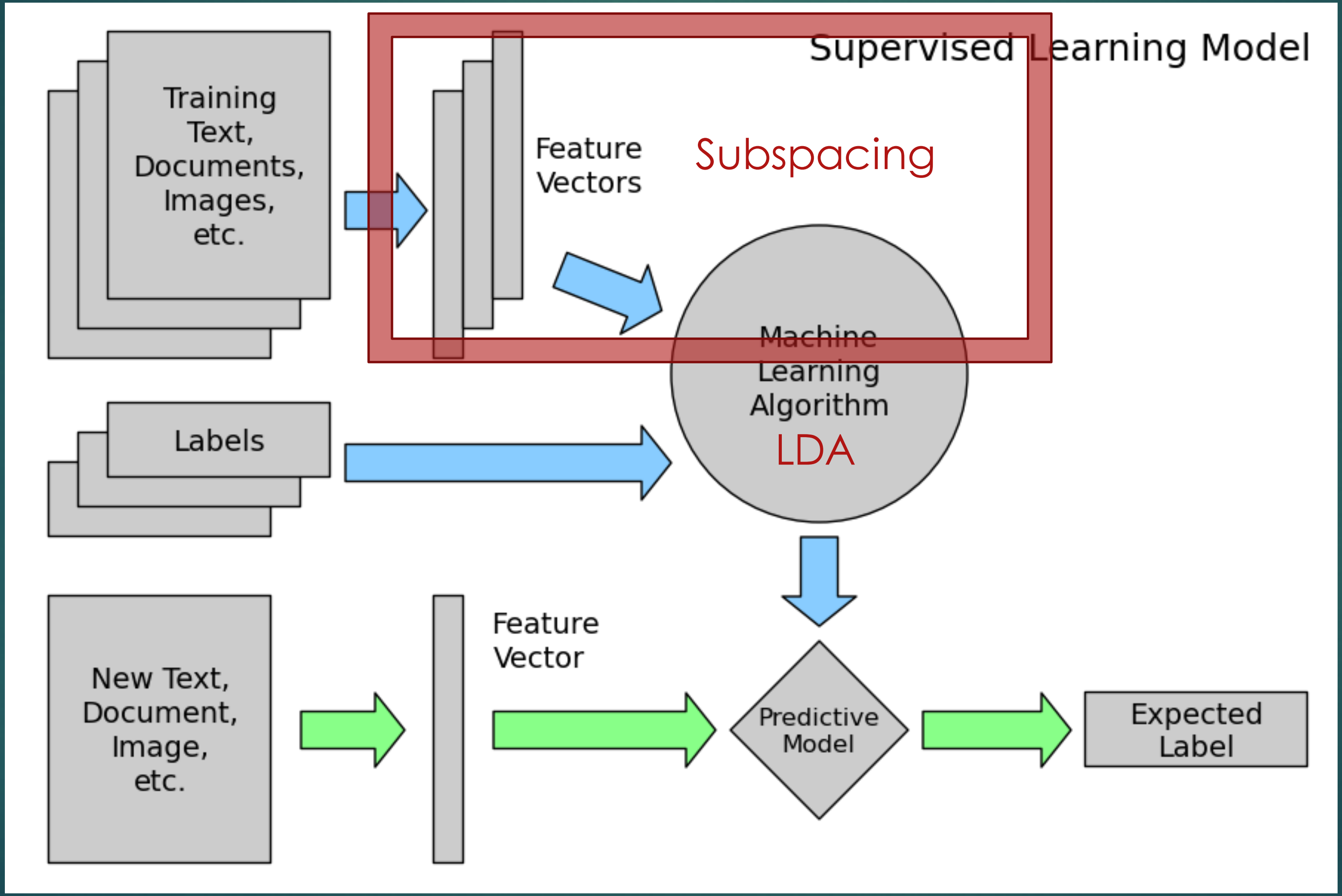
Feature set subspaceing – Efstathios Stamatatos

GIT: STAMATOS06

PRESENTED BY TIMO SOMMER

Supervised Learning Model


Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Subspacing
Machine Learning Algorithm
LDA
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

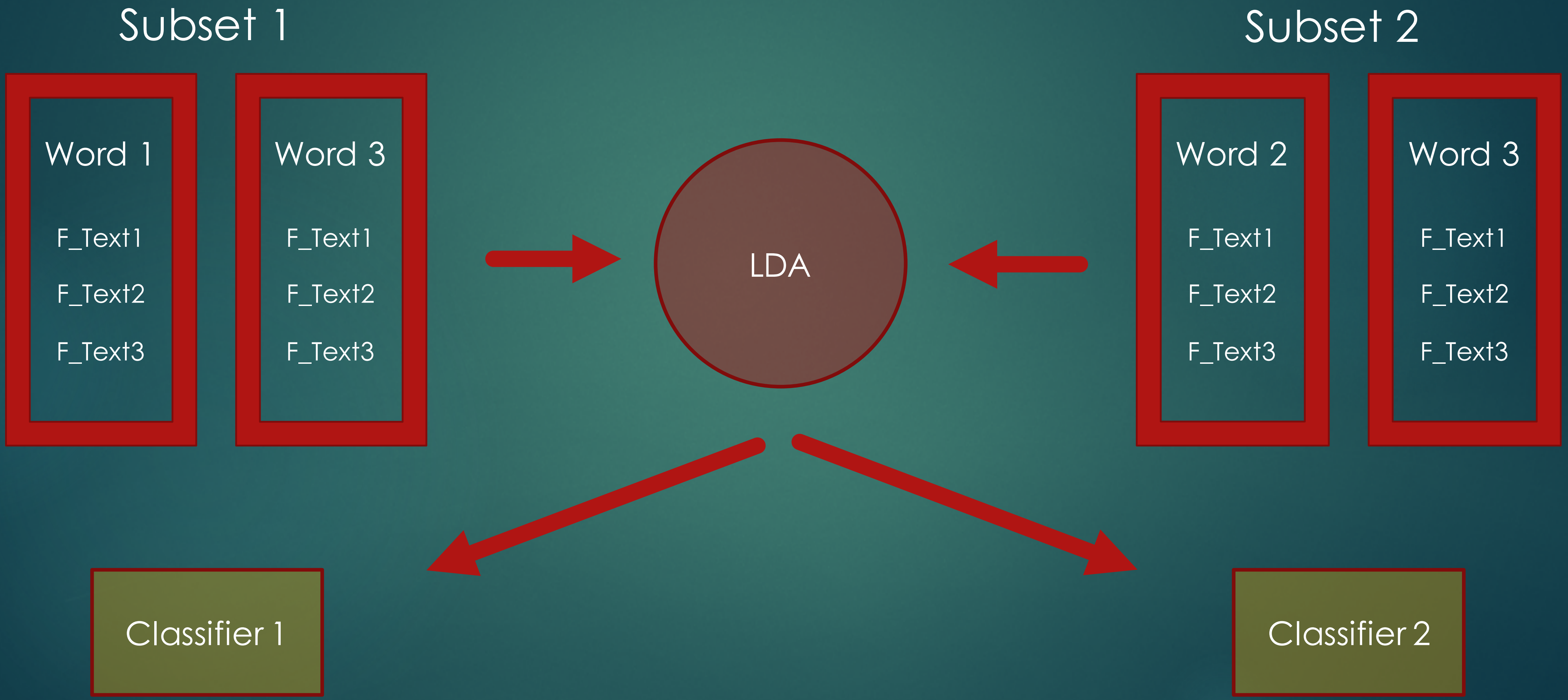
Feature set subspaceing

Word 1	Word 2	Word 3
F_Text1	F_Text1	F_Text1
F_Text2	F_Text2	F_Text2
F_Text3	F_Text3	F_Text3

Selecting Methods

- k-Random Classifier
- Exhaustive Disjoint Subspacing

Feature set subspaceing



Subset 1

Word 1

F_Text1

F_Text2

F_Text3

Word 3

F_Text1

F_Text2

F_Text3

LDA

Subset 2

Word 2

F_Text1

F_Text2

F_Text3

Word 3

F_Text1

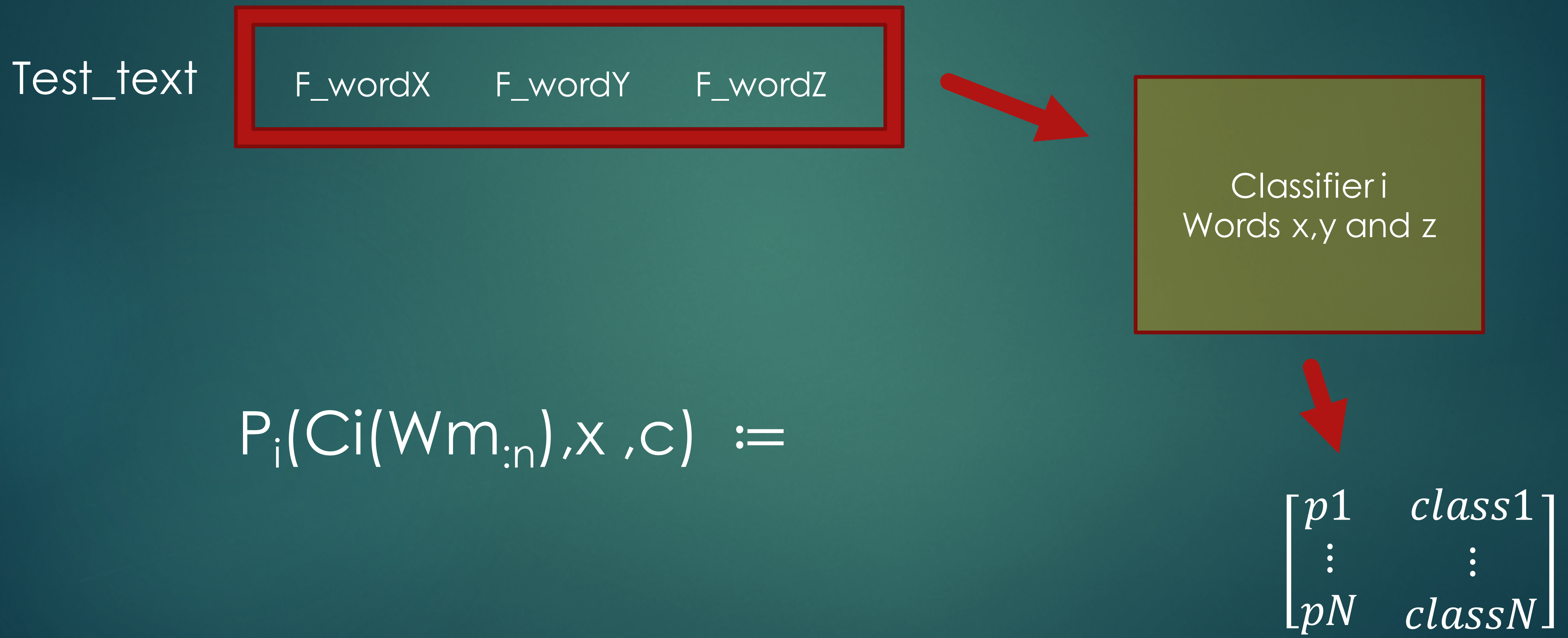
F_Text2

F_Text3

Classifier 1

Classifier 2

Posterior probabilities



Test_text

F_wordX F_wordY F_wordZ

Classifier i
Words x,y and z

$$\begin{bmatrix} p1 & class1 \\ \vdots & \vdots \\ pN & classN \end{bmatrix}$$

$P_i(Ci(Wm_{:n}),x ,c) \;:=$

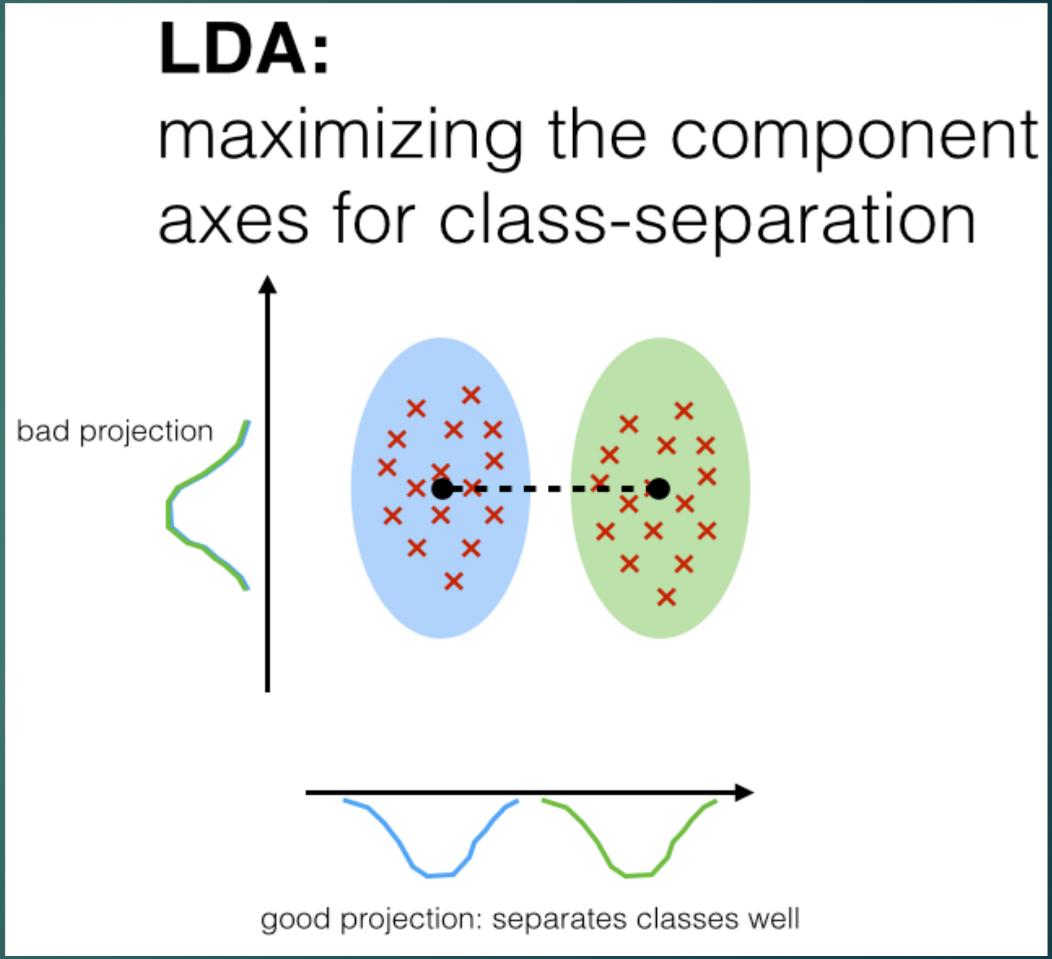
Labelling the test-text

$$\left.\begin{array}{ll}\text{Mean :} & \frac{1}{k}\sum_{i=1}^{k} \mathrm{P}_{\mathrm{i}}(\mathrm{Ci}(\mathrm{Wm}_{:\mathrm{n}}),\mathrm{x}\ ,\mathrm{c}) \\[2em] \text{Product:} & \sqrt[k]{\prod_{i=1}^{k} \mathrm{P}_{\mathrm{i}}(\mathrm{Ci}(\mathrm{Wm}_{:\mathrm{n}}),\mathrm{x}\ ,\mathrm{c})}\end{array}\right\} \text{Combined to mp (average)}$$

LINEAR DISCRIMINANT ANALYSIS

- Lots of math

- Provide posterior probabilities



LDA:

maximizing the component axes for class-separation

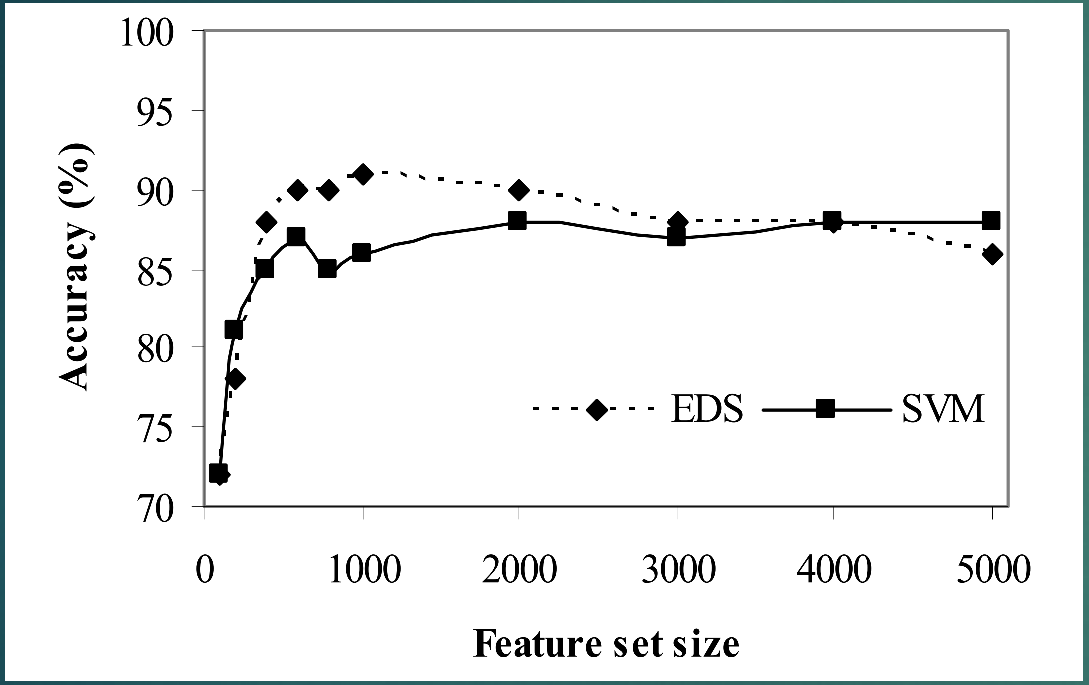
Reproduction

- Python 2.7
- Numpy
- Scikit Lern
  - Provided LDA with posterior probability
  - Provided a tokenizer for words

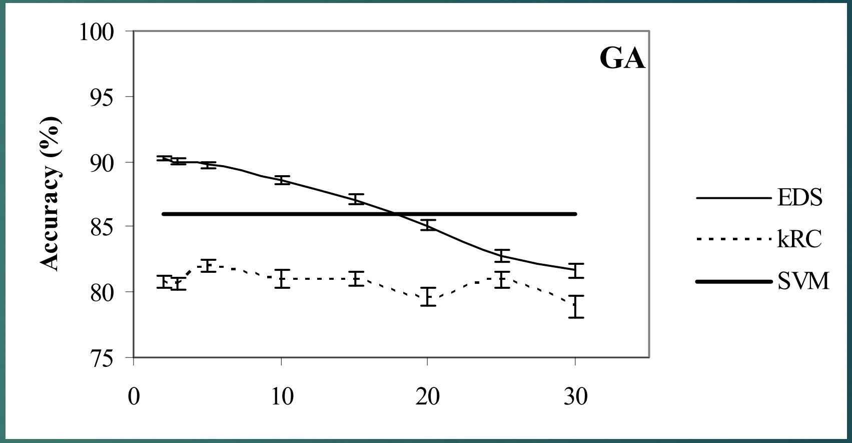
Dataset:

Vima- Dataset
Greek newspapers 2 x 10 authors with 10 training and 10 test text each
average length: 866.8 and 1148.2 words

Settings



Feature set size : 1000
Subset length : 2



Experiment results

For n = 1000, m=2

corpus		kRC Ensemble Double k	ESD Ensemble
GB	In paper	98%	99 %
GB		87%	92 %
GA	In Paper	86%	90 %
GA		75%	83%

Problems

- Finding an appropriate machine learning library
- Python: whitespace can cause errors
- Focus on the simple models not on the stacked ones

Advantages and Disadvantages

Advantages

- Language Independent
- Good performance even with text shorter then 1000 words

Disadvantages

- For large feature sets and subsets the possible feature groupings grow exponential and Training time as well
- Cannot solve the open-class Problem, occurs when the author is not in the training set
- Not independent from the number of training texts per author

Reference

Stamatatos, E. (2006). Authorship Attribution Based on Feature Set Subspacing Ensembles,
 Int. Journal on Artificial Intelligence Tools, 15(5), pp. 823-838, World Scientific
Machine Learning 101- http://www.astroml.org/sklearn_tutorial/general_concepts.html
Sebastian Raschka , http://sebastianraschka.com/Articles/2014_python_lda.html
Aly A. Farag , Shireen Y. Elhabian A Tutorial on Data Reduction (LDA)