Text Categorization

Student: Melemciuc Marius-Constantin

Problema

- Clasificarea documentelor pe topic-uri/domenii
- Dataset Reuters-21578
  - 21578 Docs
    - 20856 Train Docs
    - 722 Test Docs

- Luăm cele mai populare 10 topic-uri - T
  - Pentru fiecare topic T_i, $0 < i < 10$
    - ***Train***
      - Documentul de train D_{k1} este de tipul T sau non-T

      Avem construit **modelul**
    - ***Test***
      - Clasificăm documentul de test D_{k2} ca T sau non-T, pe baza **modelului**
    - Pe baza Test obținem
      - *Precision* - abilitatea de a nu clasifica pozitiv un doc ce este negativ
      - *Recall* - abilitatea de a clasifica corect **toate** pozitivele
      - *F1* - media armonică *Precision* și *Recall*

Exemplu Test

- Document Labels
  - **grain**
  - **non-grain**

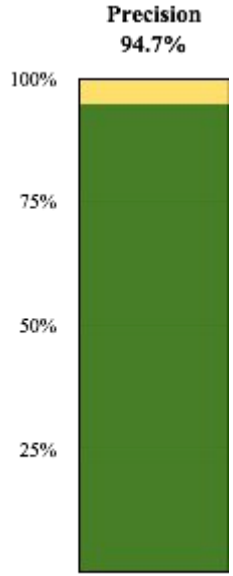
- Train Docs
  - 583 **grain**
  - 20273 **non-grain**

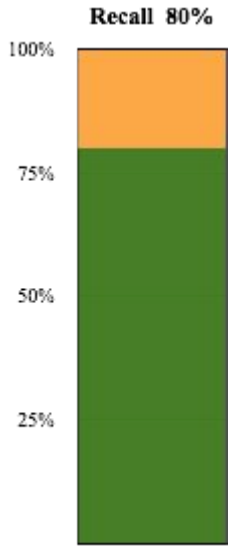
- Test Docs
  - 45 **grain**
  - 677 **non-grain**

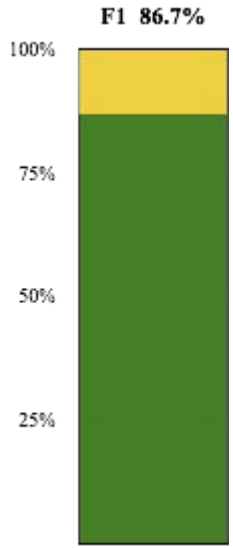
- Confusion Matrix

		Predicted Class: Yes	Predicted Class: No
Actual Class	Yes	TP	FN
	No	FP	TN

36	2
9	675



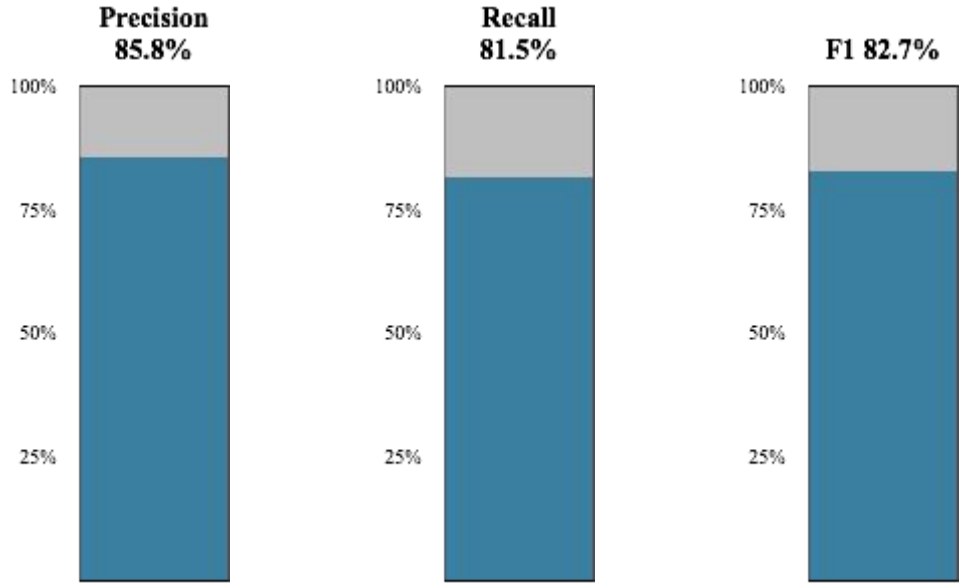




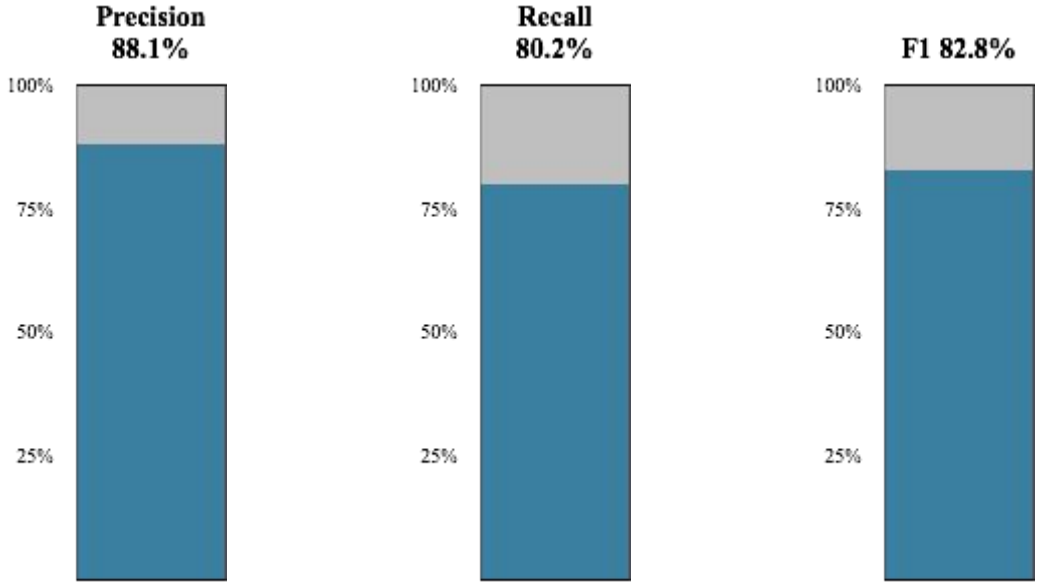
Media testelor

Support Vector Machine - SVM

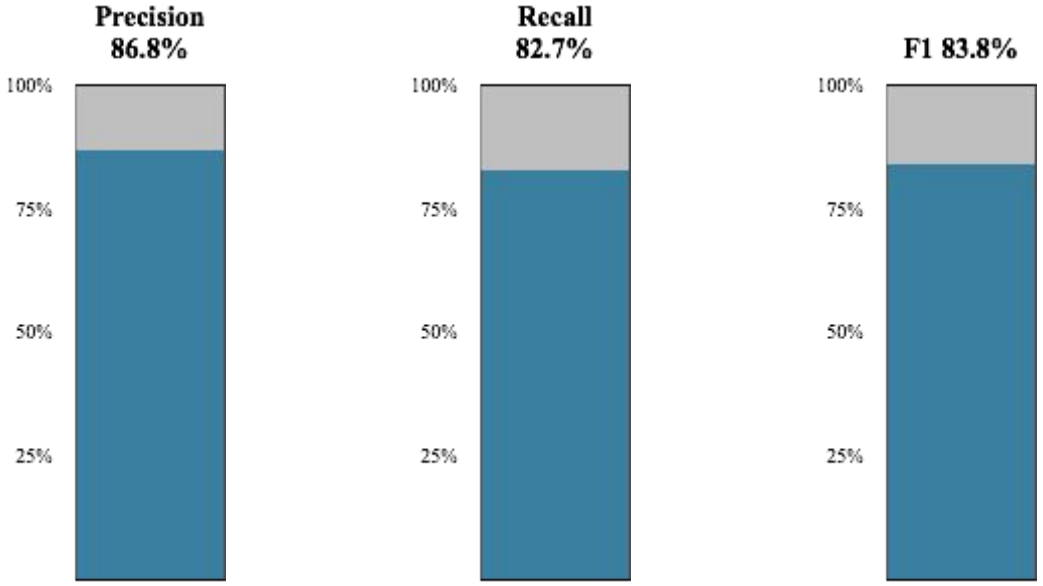
- Aplicând
  - Remove stopwords
  - Stemming



- Aplicând
  - Remove stopwords



- Fără
  - Remove stopwords
  - Stemming



Clasificator	Obs	Precision	Recall	F1
SVM	AB	0.858	0.815	0.827
	A	0.881	0.800	0.828
	-	0.868	0.827	0.838
Naive-Bayes	AB	0.791	0.780	0.781
	A	0.779	0.785	0.776
	-	0.771	0.786	0.771
Perceptron	AB	0.716	0.848	0.764
	A	0.752	0.868	0.795
	-	0.678	0.895	0.746

A = with remove stopwords, B = with stemming