
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-TAvem 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**

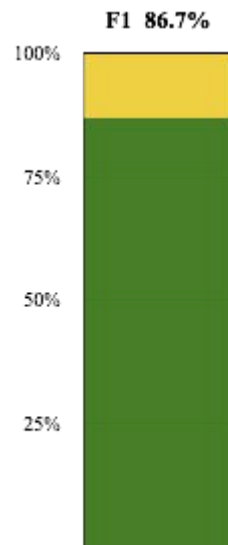
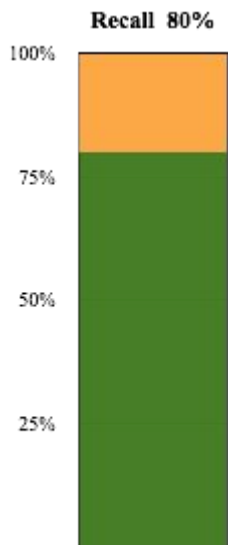
- Test Docs

- 45 **grain**
- 677 **non-grain**

- Confusion Matrix

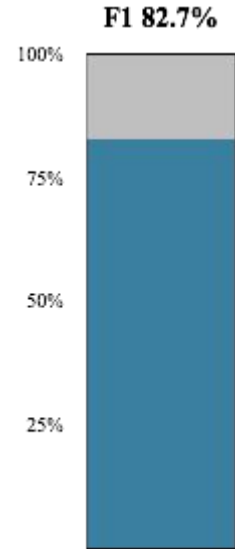
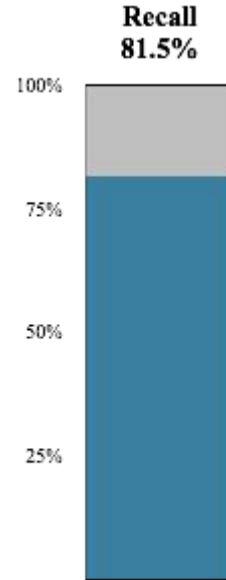
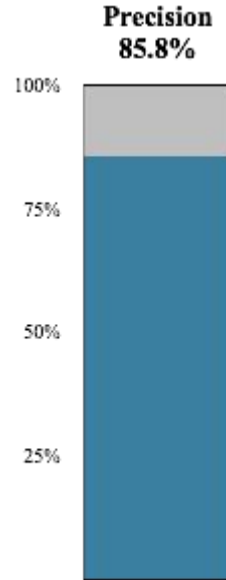
		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

36	2
9	675



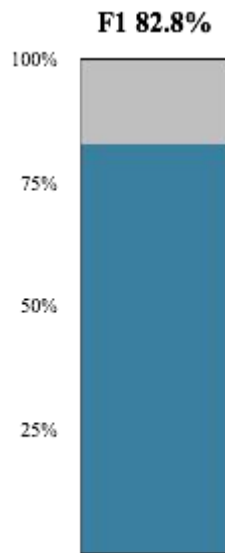
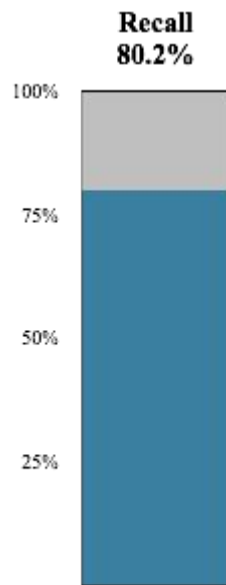
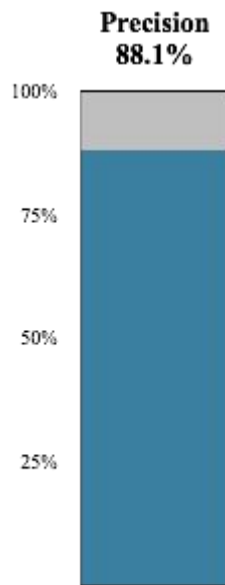
Media testelor

- Aplicând
 - Remove stopwords
 - Stemming



- Aplicând

- Remove stopwords



- Fără

- Remove stopwords
- Stemming

