Text Categorization

Student: Melemciuc Marius-Constantin

Problem

- Classify documents on topics/domains
- Dataset Reuters-21578
  - 21578 Docs
    - 20856 Train Docs
    - 722 Test Docs
- Take the 10 most popular topics
  - For every topic T_i, $0 < i < 10$
    - ***Train***
      - Train document D_{k1} has type T or non-T, but not both
        Now we have the **trained model**
    - ***Test***
      - Classify test document D_{k2} as T or non-T, based on the **training model**
    - After Test we obtain
      - *Precision* - ability of the classifier not to label as positive a sample that is negative
      - *Recall* - ability of the classifier to find **all** the positive samples
      - *F1* - harmonic mean of the *Precision* and *Recall*

Test Example

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

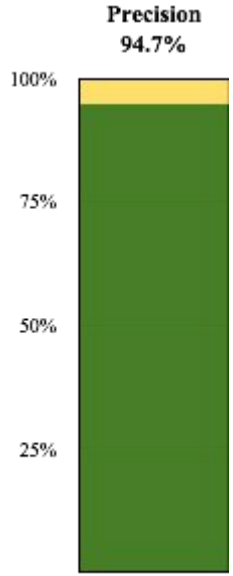
- Train Docs
  - 583 **grain**
  - 20273 **non-grain**

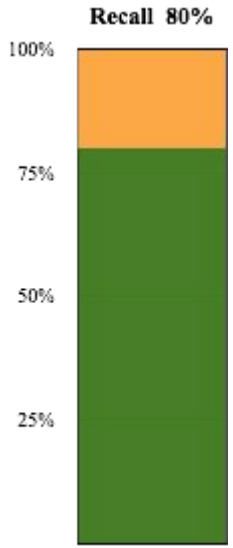
- Test Docs
  - 45 **grain**
  - 677 **non-grain**

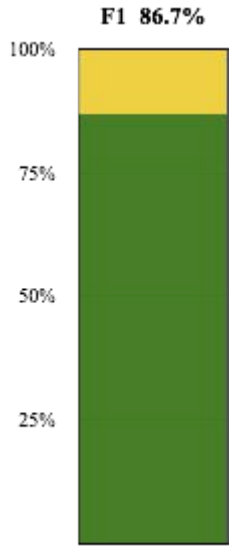
Confusion Matrix

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

36	2
9	675

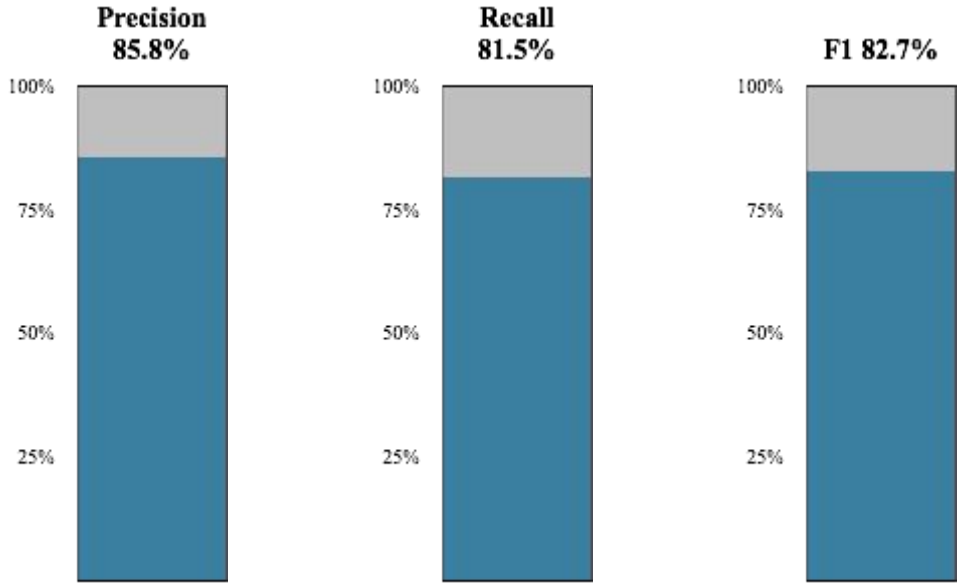




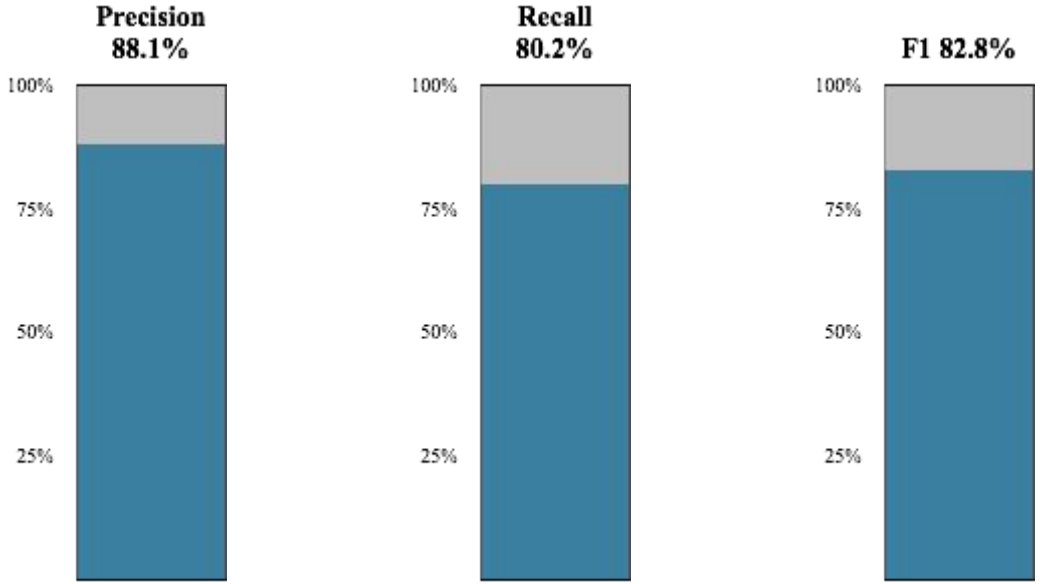


Tests Average

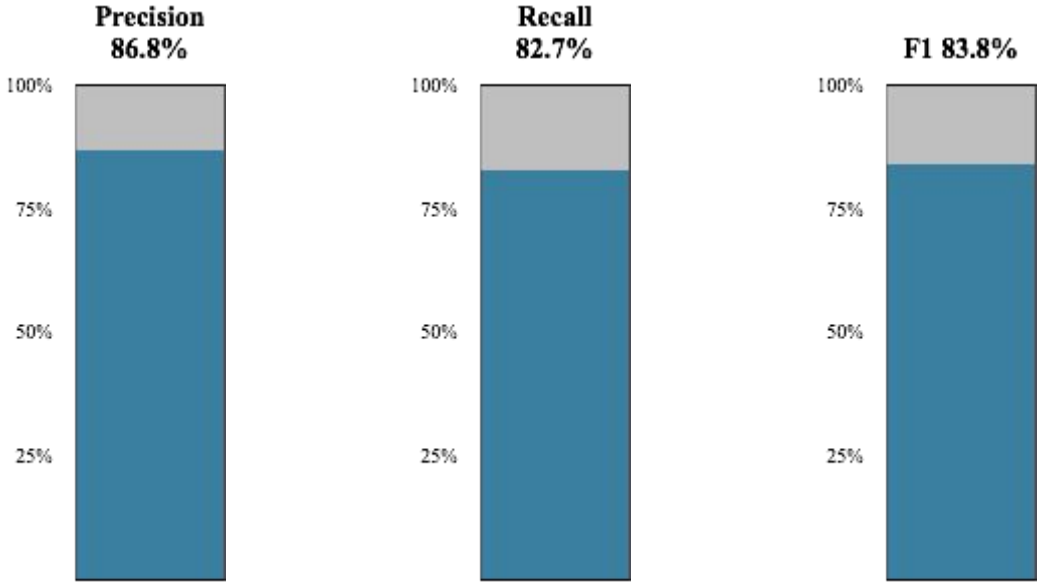
- Using
    - Remove stopwords
    - Stemming



- Using
  - Remove stopwords



- Without
  - Remove stopwords
  - Stemming



Classifier	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