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**
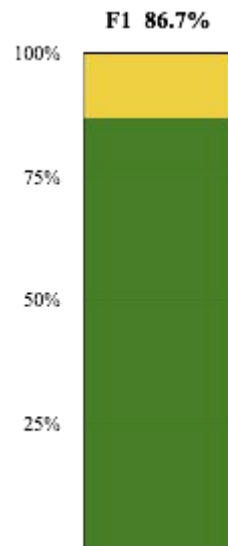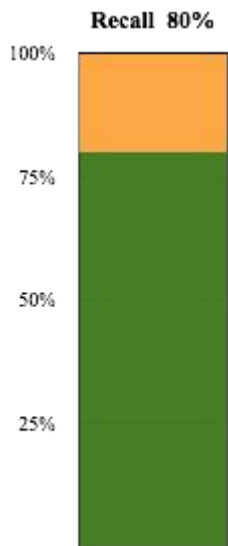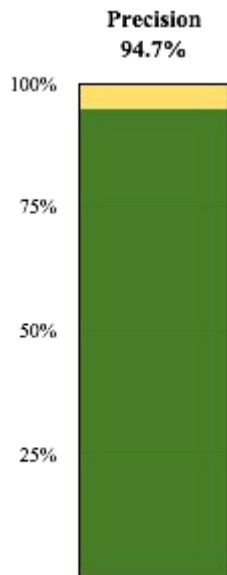
  - 20273 **non-grain**

- Test Docs

  - 45 **grain**

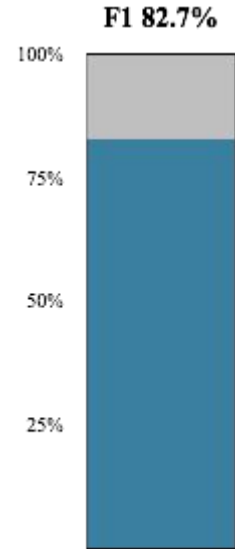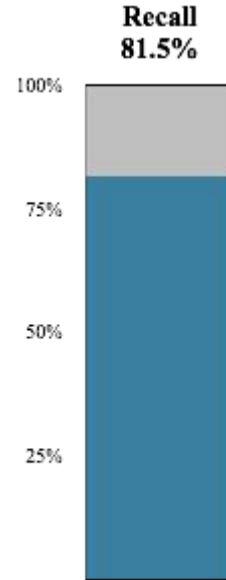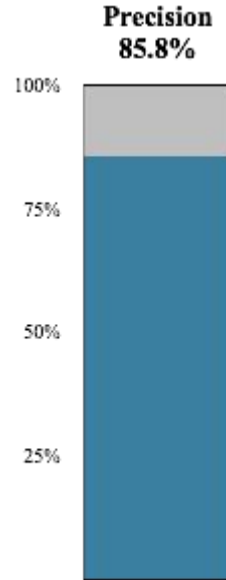 - 677 **non-grain**

- **Confusion Matrix**

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

36	2
9	675

Precision 94.7%

100%	
75%	
50%	
25%	

Recall 80%

100%	
75%	
50%	
25%	

F1 86.7%

100%	
75%	
50%	
25%	

Tests Average

- Using

  - Remove stopwords
  - Stemming

Precision
85.8%

Recall
81.5%

F1 82.7%

● Using

　　○ Remove stopwords

Precision
88.1%

Recall
80.2%

F1 82.8%
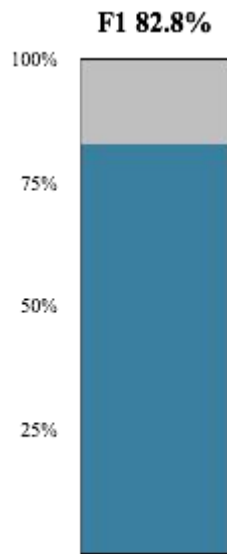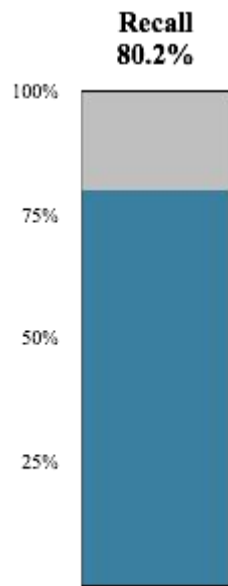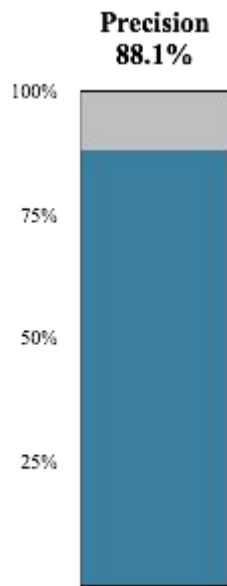
- Without

  - Remove stopwords
  - Stemming