UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

PROJECT:

TEXT MINING & SEARCH

PAOLO CAGGIANO - DAVIDE GIARDINI

IDEA

MEDICAL ABSTRACT

- MCML CLASSIFICATION
- SUMMARIZATION

DATASET

It is composed of 9445 observations describing 5 different classes of patient conditions:

- Neoplasms
- Digestive system diseases
- Nervous system diseases
- Cardiovascular diseases
- General pathological conditions

TEXT PRE-PROCESSING

- Basic preprocessing (remove punctuation, set words to lowercase, ecc.)
- Stopwords Removal
- Lemmatization

FEATURE EXTRACTION:

- Bow
- TF
- Tf-idf
- Word Embeddings(Trained & Pre-Trained)

FEATURE SELECTION:

- Rare words removal
- PCA

CLASSIFICATION ALGORITHMS

Four classification algorithms:

- Decision Tree
- Random Forest
- Naive Bayes
- SVM

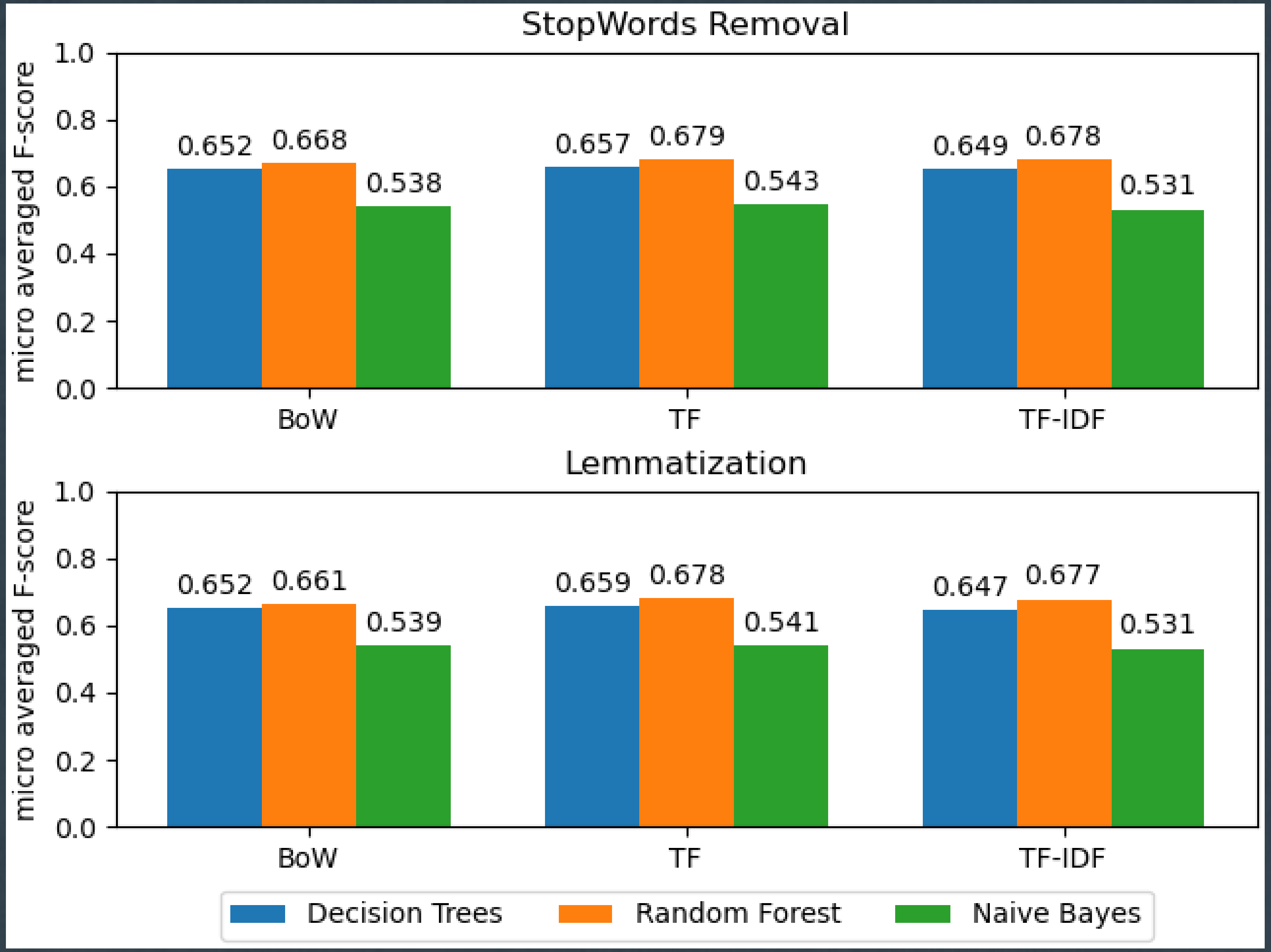
PERFORMANCE METRICS

micro-averaged:

$$\left.\begin{aligned} Recall &= \frac{TP}{TP+FN} \\[2em] Precision &= \frac{TP}{TP+FP} \end{aligned}\right\} F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

RESULTS 1/4

- Two feature extraction methods
- Two preprocessing techniques
- Three classifiers



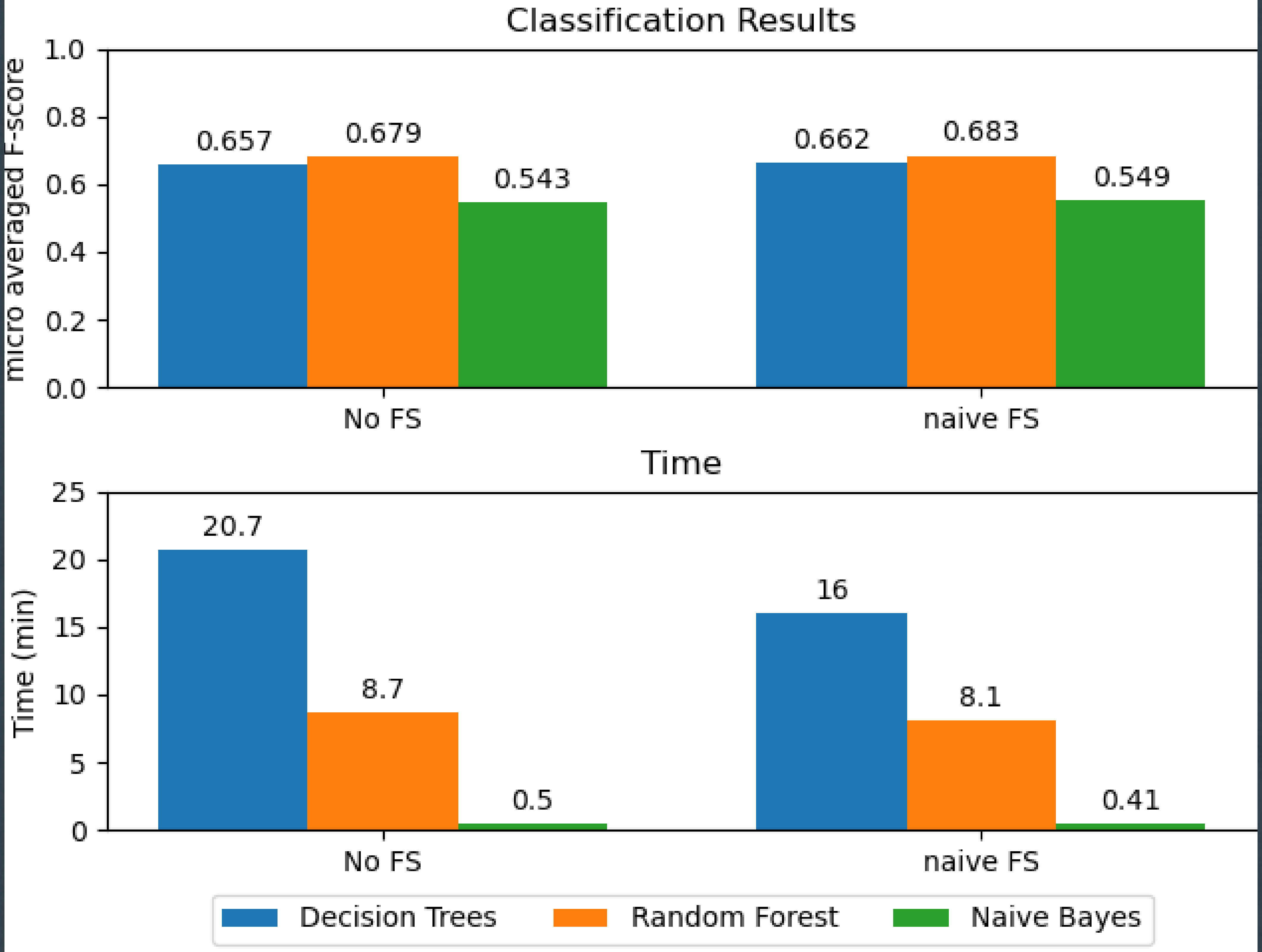
In order to get a more effective estimate of the classifier's performances we use for all of the analysis the 5 Fold cross validation

RESULTS 2/4

Naive Feature Selection

Comparison based on:

- Performance
- Execution time



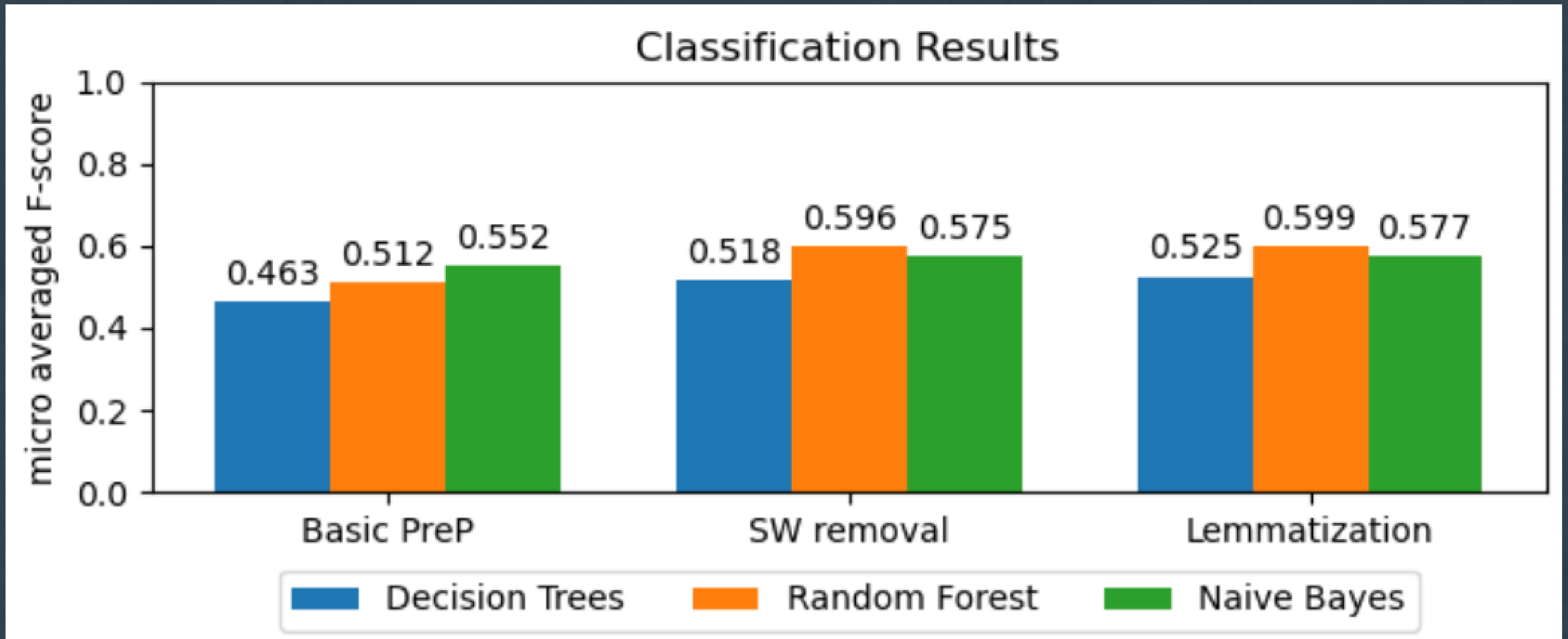
RESULTS 3/4

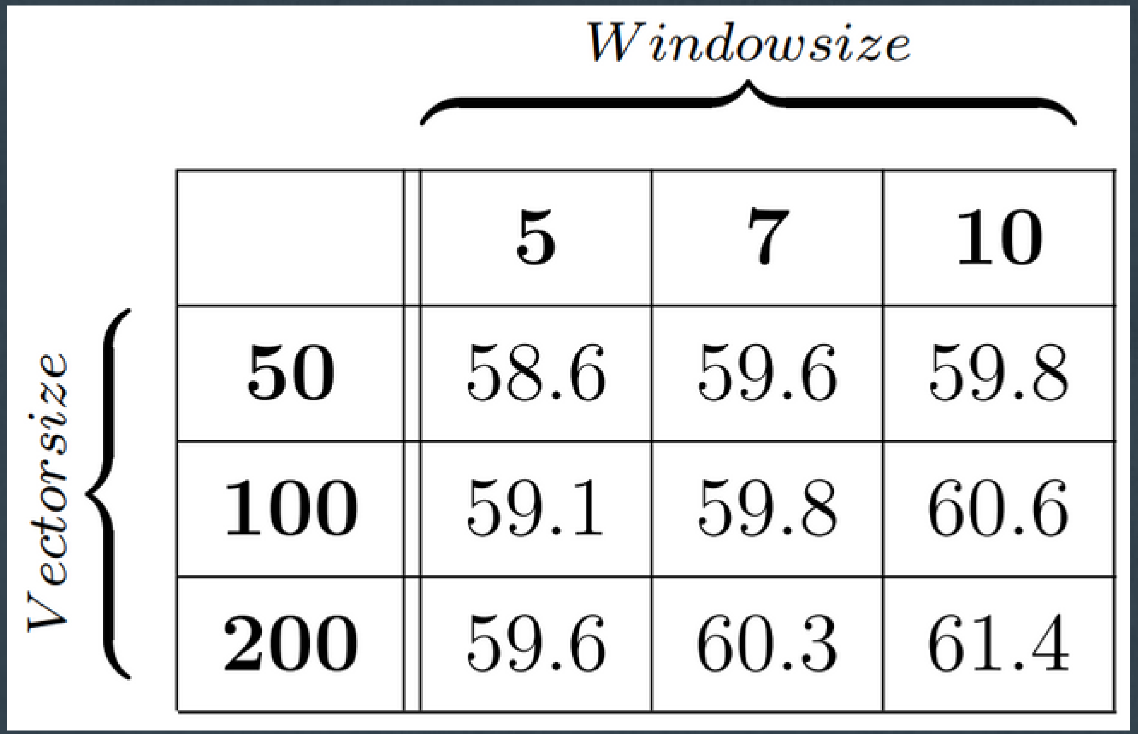
Training Word Embeddings

word2vec model:

- vector size = 100
- window = 7

We test different combinations:





		Windowsize		
		5	**7**	**10**
Vectorsize	**50**	58.6	59.6	59.8
	100	59.1	59.8	60.6
	200	59.6	60.3	61.4

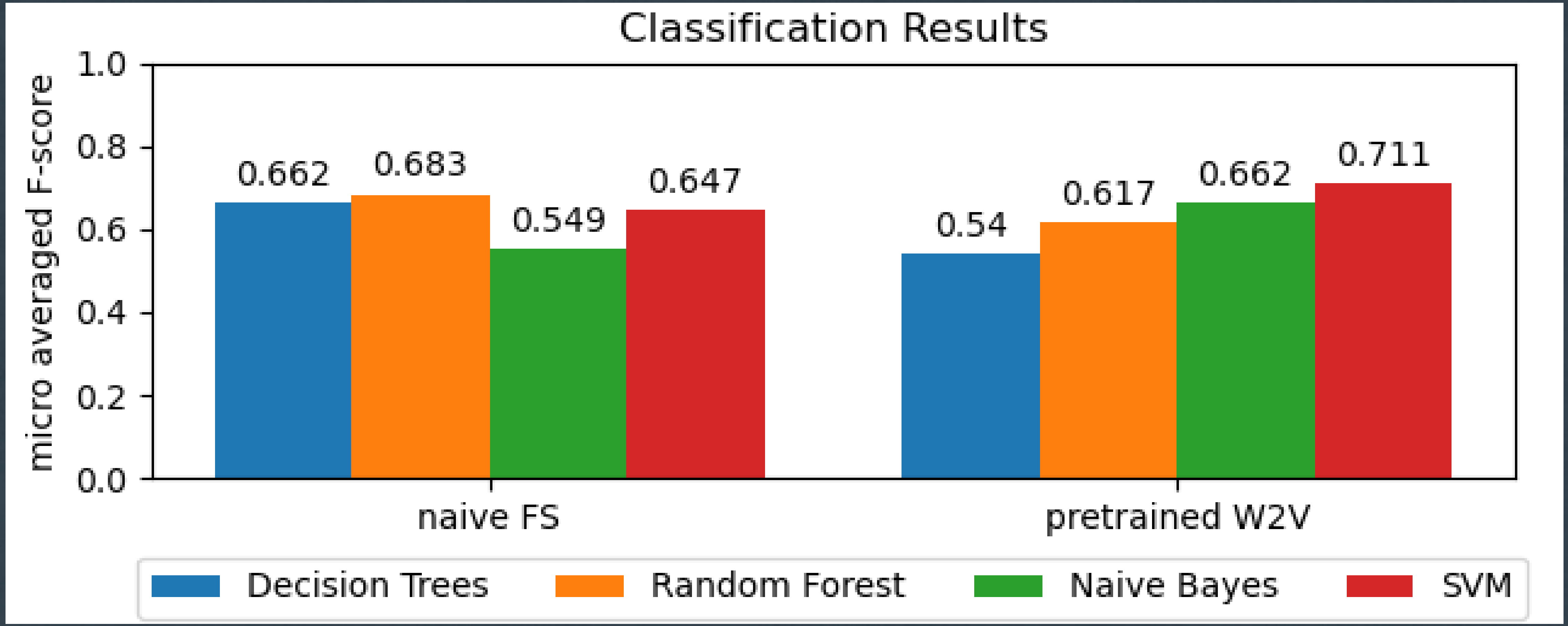
RESULTS 4/4

Pre-Trained Word embeddings

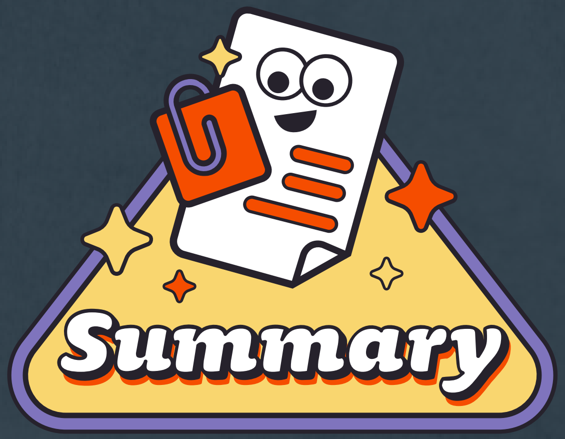
Best combination so far:
(StopWords Removal + TF + Feature Selection)

vs

Pretrained word embeddings:
from a combination of PubMed and PMC



SUMMARIZATION



- ABSTRACT

expresses the ideas in the source documents using different words.

- EXTRACTIVE

the summary is created from important phrases or sentences selected from the input text.

  - Graph-based method

    Represent the document as a connected graph where vertices are the sentences, and edges reflect their similarity.
    After we use Page Rank Algorithm to retrieve the score of each sentence

  - LSA

    it is an algebraic-statistical method that extracts hidden semantic structures of words and sentences.

EVALUATION

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
is a measure to automatically determine the quality of a summary
by comparing it to other (ideal) summaries created by humans

Two types used:

- Rouge-n

Computed as the number of common n grams between the candidate and reference summaries and the total number of n -grams present in the reference summary.

- Rouge-L

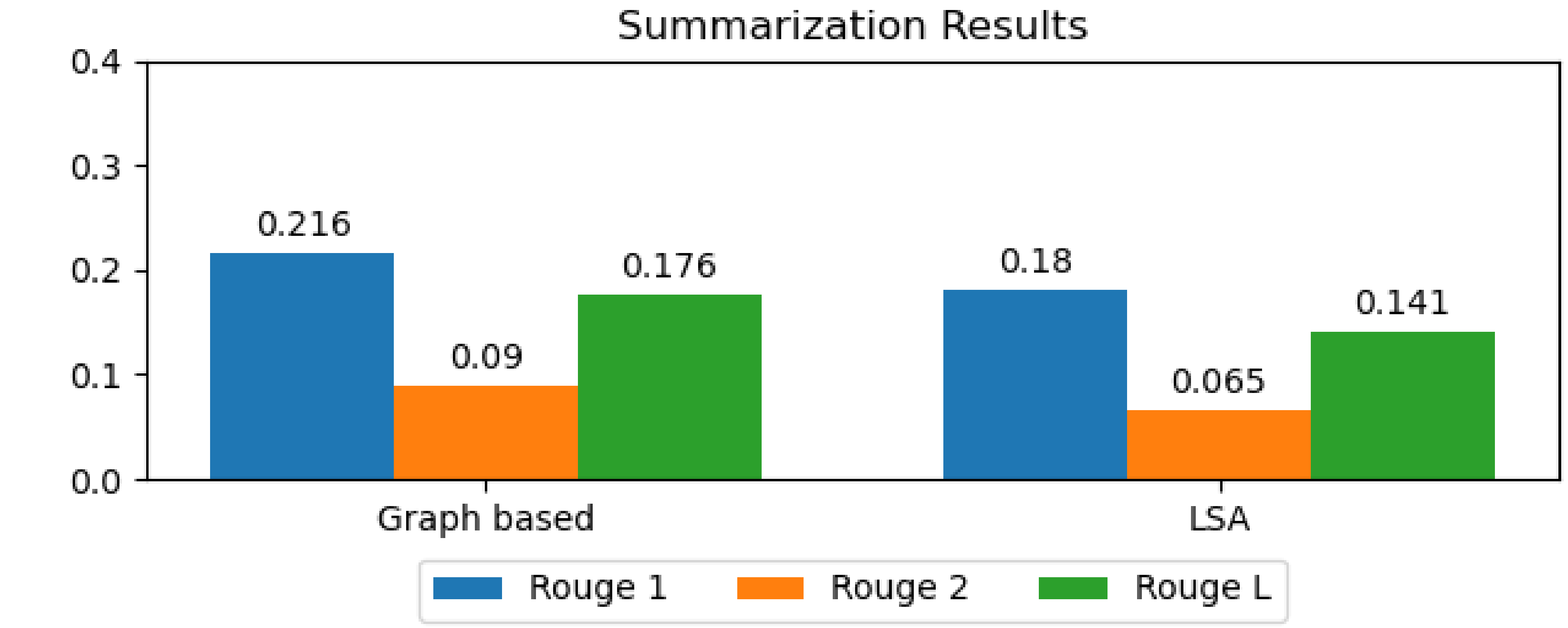
It refers to the longest common subsequence between two texts. All n-grams must be consecutive.

SUMMARIZATION RESULTS 1/2

Our dataset does not provide humans generated summary. To overcome this issue, we use document's title as reference

↓

We retrieve them with PubMed"s API



We compute the summarization only on the abstract part of the observation, and we compare it to the title via the ROUGE metric.

SUMMARIZATION RESULTS 2/2

We determine how the summaries perform in the previous classification task. More precisely, we use the pretrained word embeddings to perform feature extraction over the two summaries (Graph based and LSA).

Then, we compare their results with the same classification done on the original text, and on a random summary.

DATA	F-score SVM
Entire Document	71.1%
Random Summary	63.3%
Graph Model	64.0%
LSA	62.0%

THANKS FOR THE ATTENTION