Qatar Content Classification

Client

Tarek Kanan

tarekk@vt.edu

Presenter

Mohamed Handosa

handosa@vt.edu

VT, CS6604

May 6, 2014

About The Project

- Funded by QNRF (http://elisq.qu.edu.qa)
- Started at VT in 1/1/2013, and running through 12/31/2015.
- A project to advance digital libraries in the country of Qatar.
- Collaborating institutes: Penn State, Texas A&M, and Qatar University.

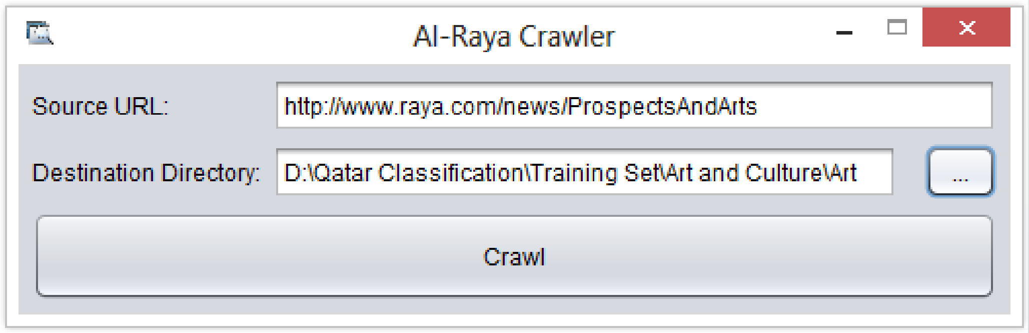
Project Plan

- Build Arabic collections using Heritrix crawler
- Build a universal taxonomy for Arabic newspapers
- Use different classifiers to classify Arabic documents
- Use Apache Solr to index and search Arabic collections
- Evaluate the performance of the classifiers on Arabic data

Building Arabic Collections

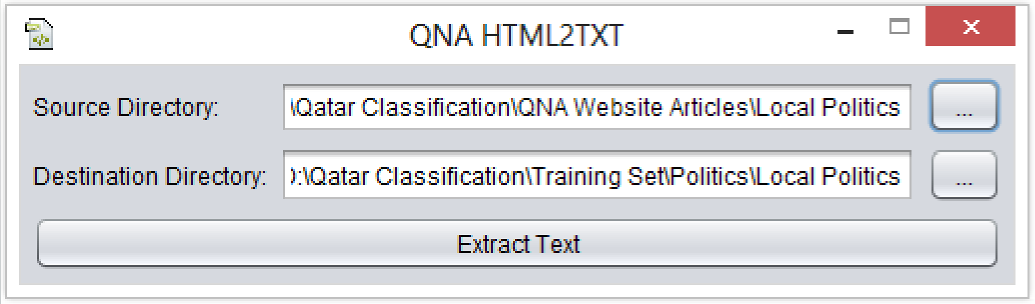
Data Sources

- Al-Raya newspaper
  - 300 online articles using “Al-Raya Crawler”.
  - 1700 newspaper PDFs using “Heritrix Crawler”.
- Qatar News Agency (QNA)
  - 450 online articles using “QNA HTML2TXT”.



Source URL: http://www.raya.com/news/ProspectsAndArts

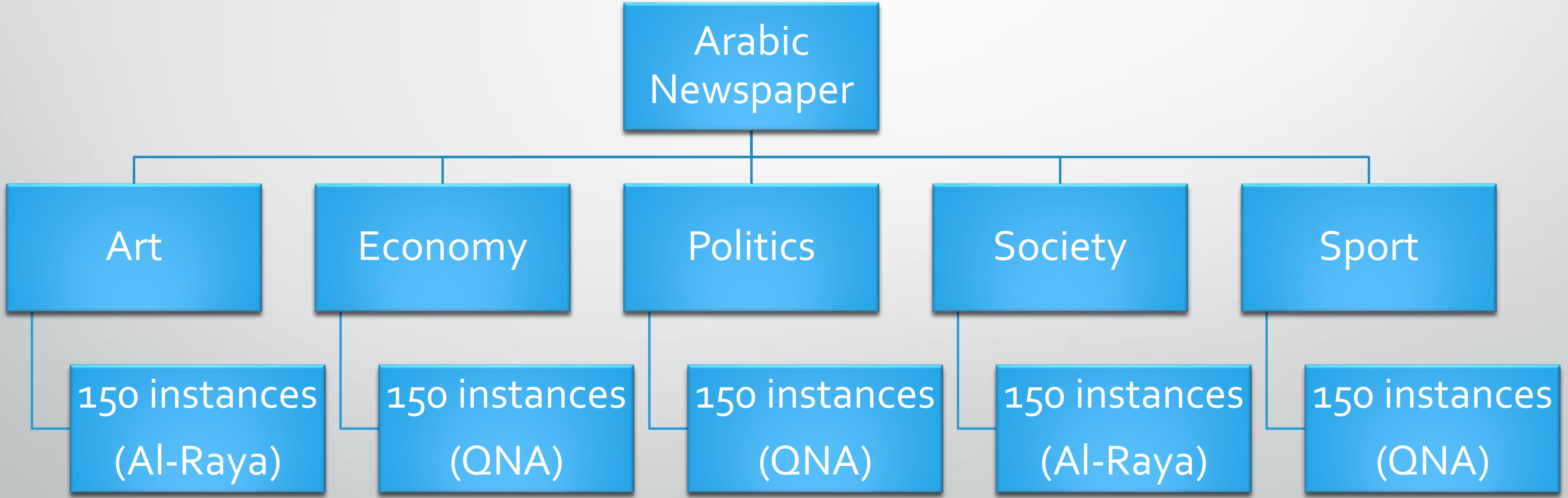
Destination Directory: D:\Qatar Classification\Training Set\Art and Culture\Art



Source Directory: \Qatar Classification\QNA Website Articles\Local Politics

Destination Directory:):\Qatar Classification\Training Set\Politics\Local Politics

Arabic Newspapers Taxonomy



Arabic Newspaper

Art

150 instances (Al-Raya)

Economy

150 instances (QNA)

Politics

150 instances (QNA)

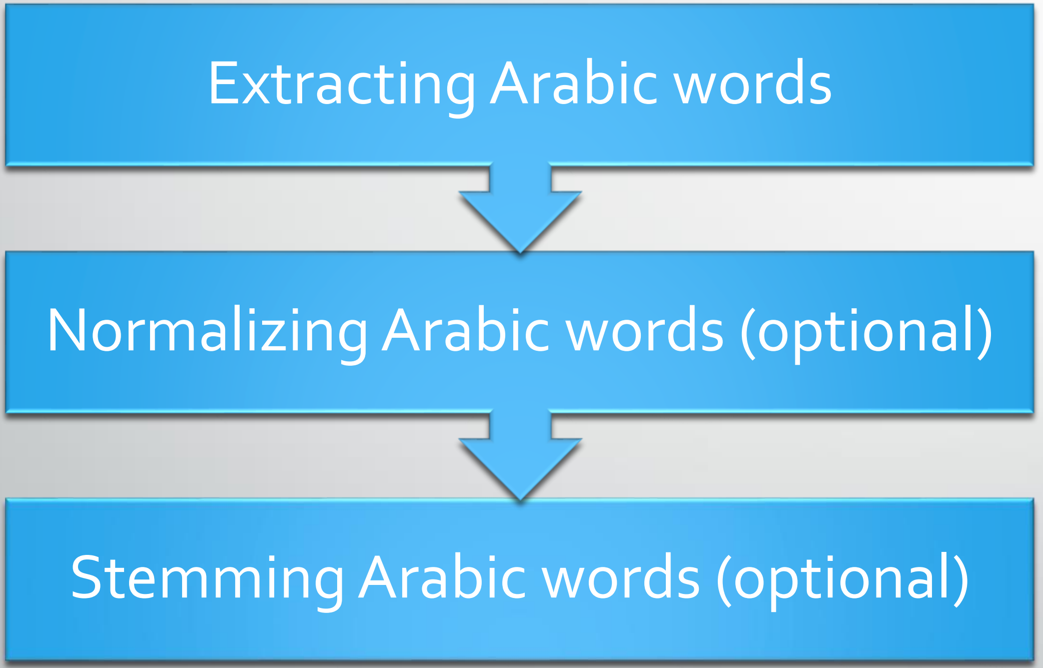
Society

150 instances (Al-Raya)

Sport

150 instances (QNA)

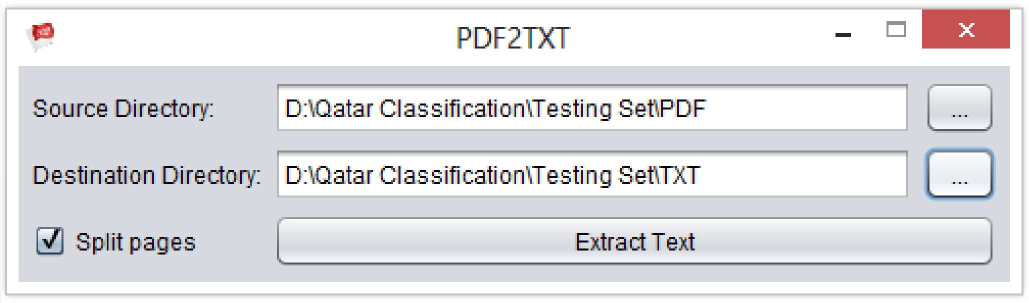
Collection Preprocessing



Extracting Arabic words

Normalizing Arabic words (optional)

Stemming Arabic words (optional)

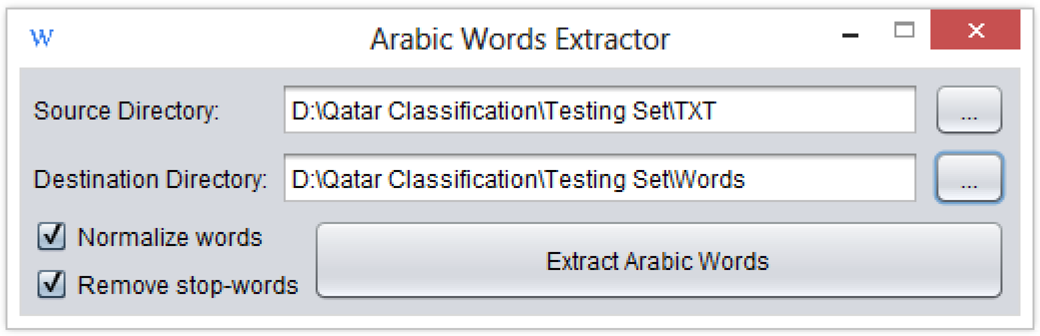


PDF2TXT

Source Directory: D:\Qatar Classification\Testing Set\PDF

Destination Directory: D:\Qatar Classification\Testing Set\TXT

☑ Split pages



Arabic Words Extractor

Source Directory: D:\Qatar Classification\Testing Set\TXT

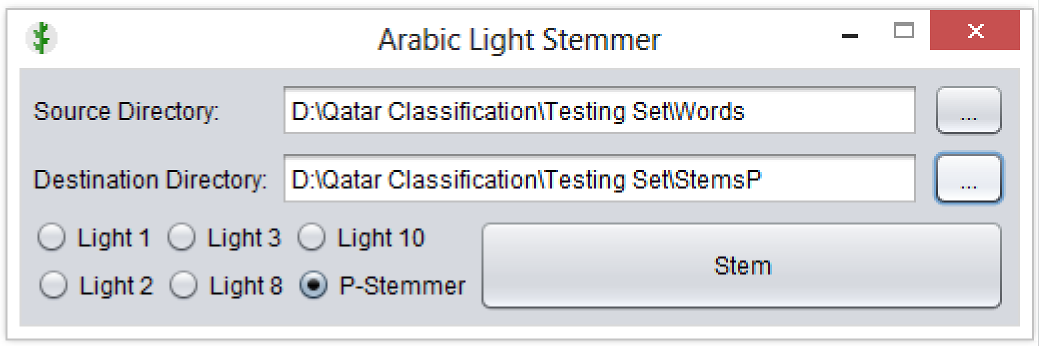
Destination Directory: D:\Qatar Classification\Testing Set\Words

☑ Normalize words

☑ Remove stop-words

Stemming Arabic Words

- Root Stemmers – too abstract.
  - “مقصد”, “الاقتصاد” → “قصد”
- Light Stemmers – widely used.
  - “مباحث”, “المباحثات” → “مباحث”
- P-Stemmer – even better ☺.
  - “مباحث” → “مباحث”
  - “المباحثات” → “مباحثات”



Arabic Light Stemmer

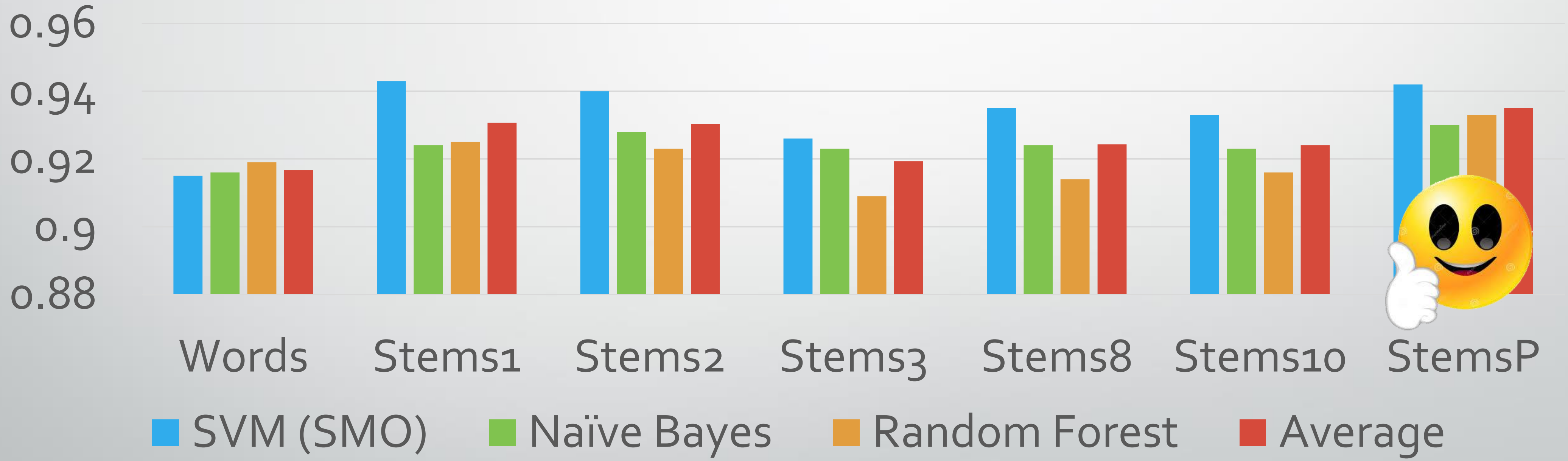
Source Directory: D:\Qatar Classification\Testing Set\Words

Destination Directory: D:\Qatar Classification\Testing Set\StemsP

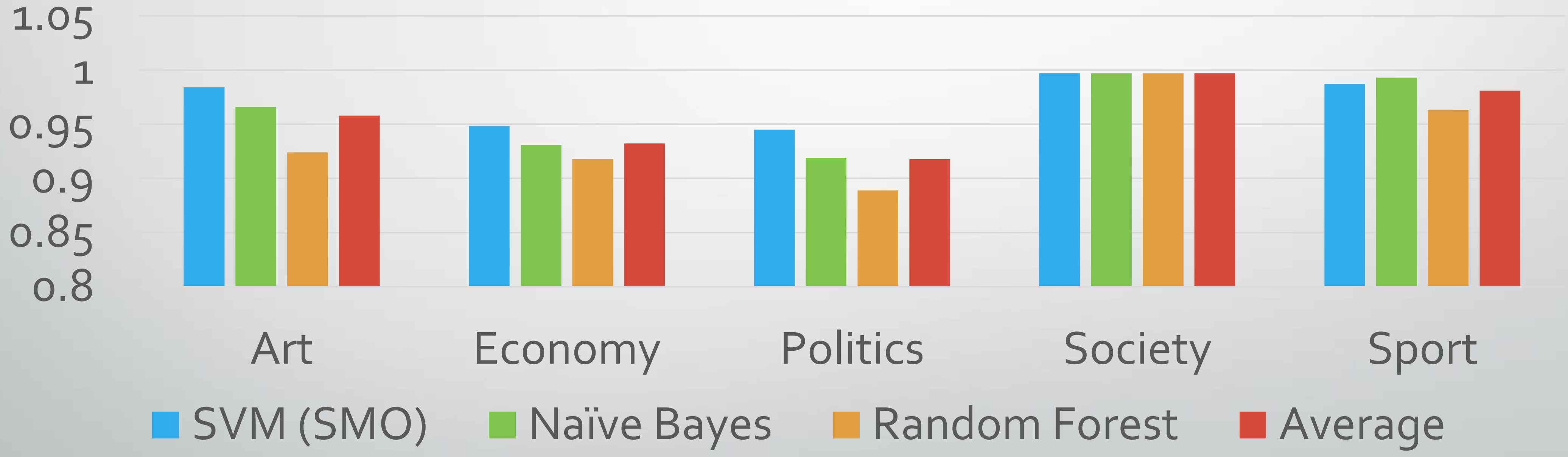
○ Light 1 ○ Light 3 ○ Light 10

○ Light 2 ○ Light 8 ● P-Stemmer

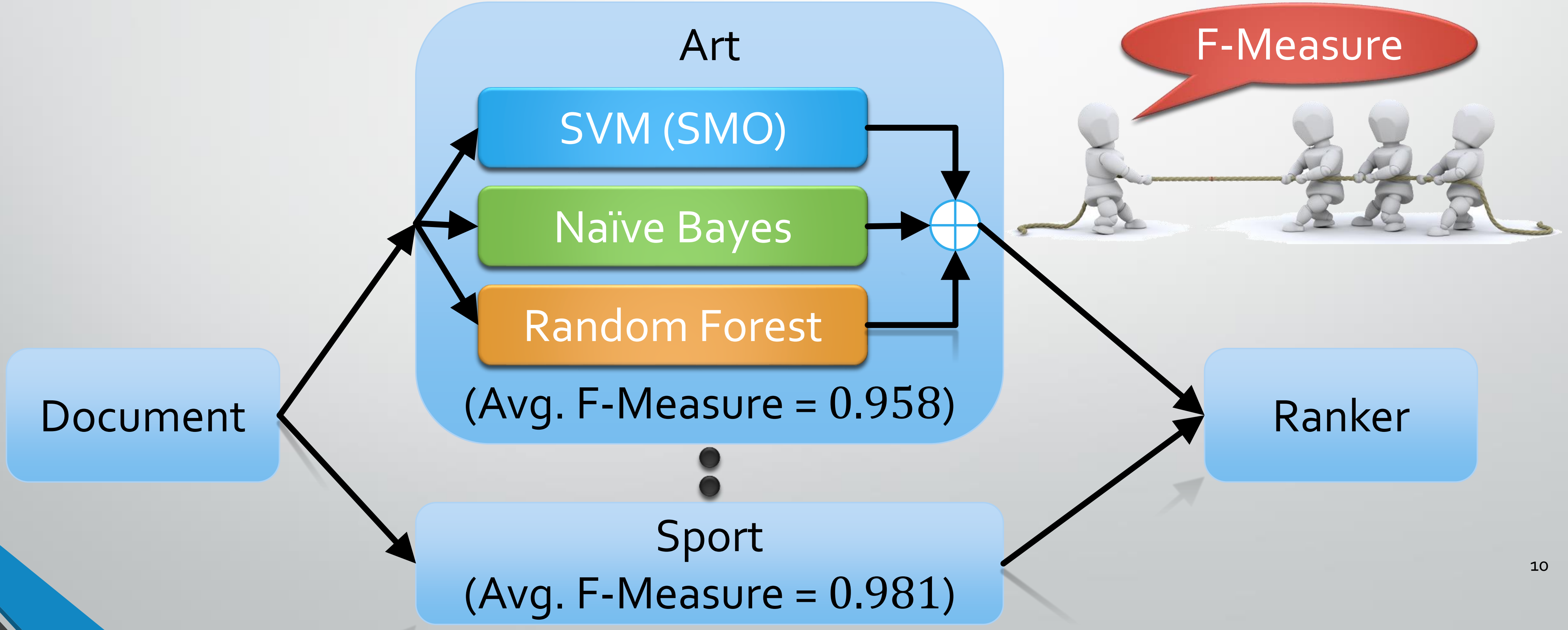
Multiclass Classification (21 Classifiers)



Binary Classification (15 Classifiers)



Arabic Newspapers Classification Framework



Art

SVM (SMO)

Naïve Bayes

Random Forest

(Avg. F-Measure = 0.958)

F-Measure

Document

Ranker

Sport

(Avg. F-Measure = 0.981)

Uploading Collections to Apache Solr

1700 Newspaper PDF Files

- Splitted into PDF text
- Converted to clean Arabic text
- Stemmed using proposed P-Stemmer.
- Classified using 5 SVM binary classifiers.
- Classified using a SVM multiclass classifier.

Solr Cores

- 11 Solr cores were created
- For each of the five binary classifiers
  - Positive instances were uploaded to a core
- For the multiclass classifier
  - Instances of each class were uploaded to a core
- All instances were uploaded to the last core

Contributions

- Building a collection for Arabic newspapers.
- Developing a set of tools to process Arabic text.
- Proposing P-Stemmer, an Arabic light stemmer.
- Comparing different text classification techniques.
- Proposing a framework for Arabic Newspapers Classification.
- Creating 11 Solr cores, 2 per class and 1 containing all instances.

Future Work

Prepare Testing Set using Solr

- Upload all instances to a Solr core
- Execute a query related to a given class
- Label search outputs as belonging to that class
- Use the labeled instances to test the classifiers

Evaluate Classifiers using Solr

- Classify all instances using a classifier X
- Upload labeled instances to a Solr core
- Execute a query related to a given class C
  - Precision: ratio of retrieved C instances to the number of retrieved instances
  - Recall: ratio of retrieved C instances to the total number of C instances in the collection

Thank You