CS6604 Course Project
Fall 2019

Automatic Classification of Arabic ETDs

Eman Abdelrahman and Fatimah Alotaibi
Supervised by: Dr. Edward Fox
December 10th, 2019
Virginia Tech, Blacksburg, 24061

Acknowledgements:

- We would like to deeply thank Dr. Fox for his continuous support.
- Also, our colleague Palakh Jude for the guidelines and assistance she provided us.
- We would also like to thank our colleague Bill Ingram for adding us to his ARC allocation.
- Special thanks to Saudi Digital Libraries for giving an account for Fatimah Alotaibi which made this project possible.
- Thanks to Institute of Museum and Library Services IMLS LG-37-19-0078-19.

Outline:

- Motivation.
- NLP in Arabic language.
- Related work.
- Dataset.
- Preprocessing.
- Experiment and results.
- Insights and future work.

Motivation

- **ETDs** are becoming the new genre.
- They need **classification** for better browsing and accessibility.
- Increasing number of universities are requesting their graduate students to deposit an **Arabic** translated version of their ETD or at least for the title and abstract.
- No prior machine learning research has been done on Arabic ETDs due to:
  - Data availability.
  - Complexity of Arabic Language.

NLP in Arabic Language

According to “Introduction to Arabic Natural Language Processing” book, Nizar Y. Habash, Morgan & Claypool Publishers, 2010:

- Vast majority of **Arab**ic **words** are morphologically complex.
- Arabic is high **inflectional** and **derivational** language.
- Arabic language has rich and complex grammatical structures.

Significant challenges to many Natural Language Processing (NLP) applications.

Related Work

Classification models performance comparison:

Paper	Classifiers	Dataset	Preprocessing	Results
Gharib, T., Habib,M., and Fayed,Z., "Arabic Text Classification Using Support Vector Machines".	• Support Vector Machine • Naive Bayes • K-Nearest Neighbor • Rocchio	1,132 documents	• Stop word removal • Stemming • Document indexing • Term selection	The SVM classifier outperforms the others when the number of features is large.
"An Intelligent System for Arabic Text" M. M. Syiam, Z. T. Fayed & M. B. Habib.	• Rocchio • K-nearest neighbor	1,132 documents	Hybrid method of statistical and light stemmers	Rocchio classifier shows 98% for the accuracy

Related Work Cont.

Building new system, comparison with other existing systems

Paper	Classifiers	Dataset	Preprocessing	Result
Alaa M. El-Halees. "Arabic Text Classification Using Maximum Entropy".	Maximum entropy method to build ArabCat system	www.aljazeera.net	Stop word removal Tokenizing stemming Part of speech	ArabCat System shows 80.48 ,80.34, 80.41 for recall, precision, and F-measure respectively, where other existing systems such as Sakhr's Categorizer show 73.78 47.35 57.68

Related Work Cont.

Classification with no preprocessing

Paper	Classifiers	Dataset	Preprocessing	Result
Sawaf, H., J Zaplo,J., and Ney, H., "Statistical Classification Methods for Arabic News Articles"	statistical methods (maximum entropy text classification)	Arabic NEWSWIRE corpus contains 33k documents	No morphological analysis.	shows 89.5, 31.5 46.61 for recall, precision, and F-measure respectively,

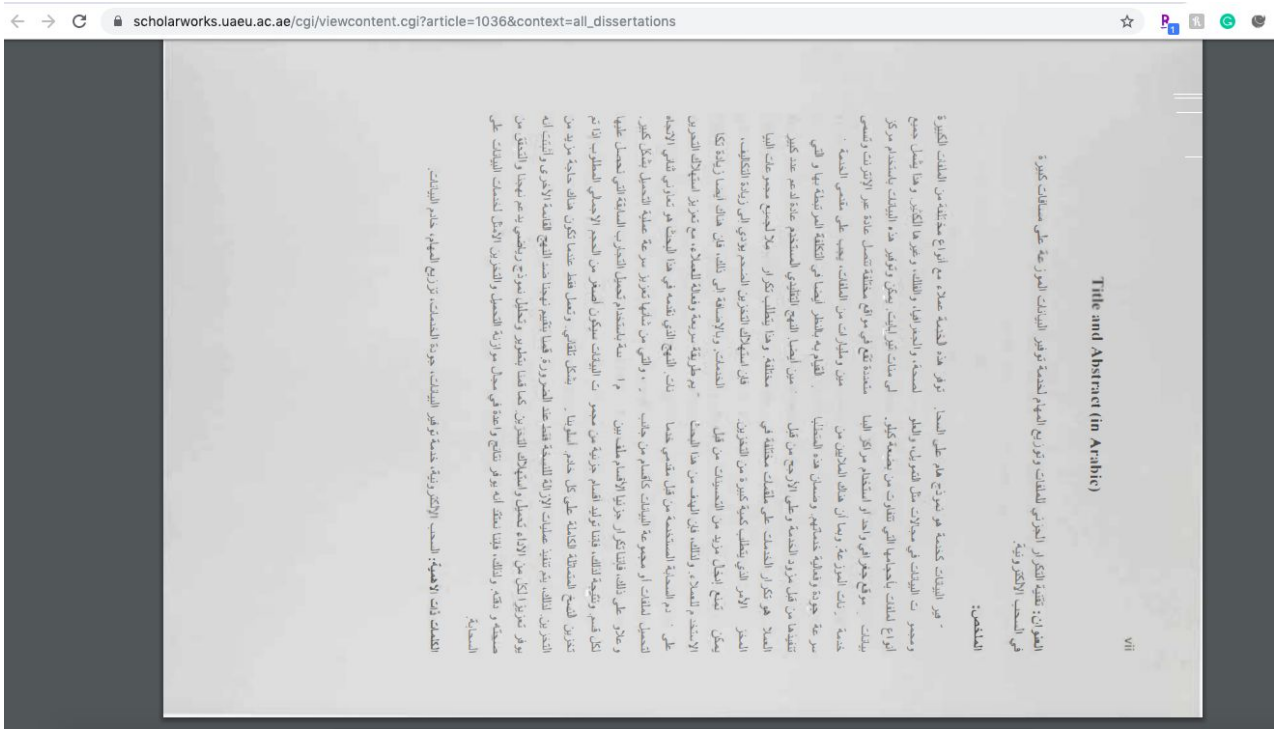
Dataset:

Dataset:

- United Arab Emirates University “Scholarworks @ UAEU”.

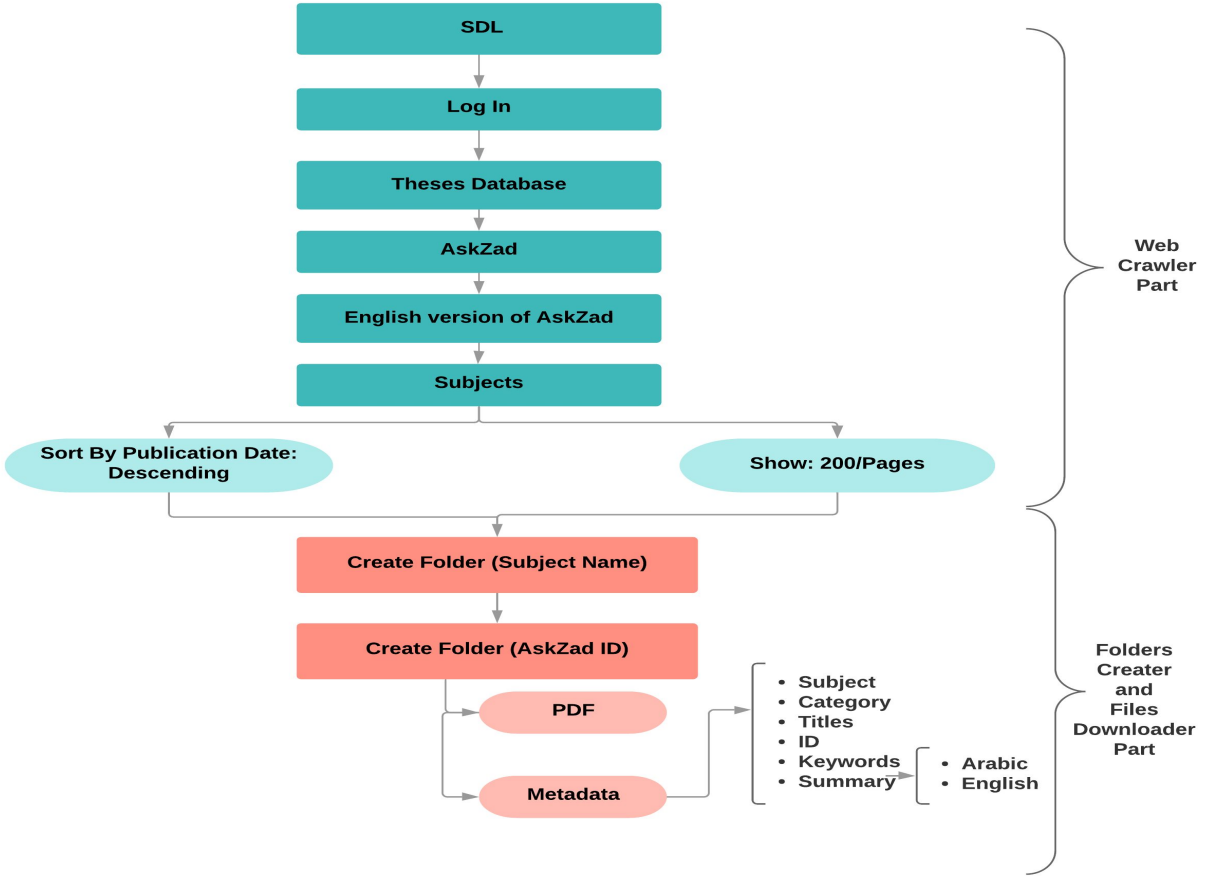
Dataset:

- United Arab Emirates University “Scholarworks @ UAEU”.
- Challenge:



Dataset

- Saudi Digital Library
  - AskZad Library



Dataset:

- Saudi Digital Library
  - AskZad Library
- Challenge:

Here are a few problems encountered while undergoing Data Extraction at a large scale:

- **Data** warehousing.
- Website Structure Changes.
- Anti- **Scraping** Technologies.
- Hostile environment/Technology.
- Honeypot traps.
- Quality of **data**.

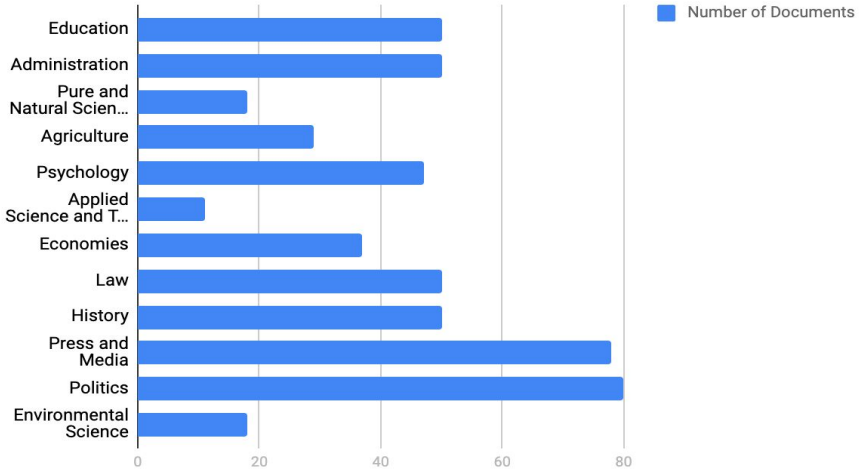
May 2, 2019

Web Scraping at Large: Data Extraction Challenges You Must ...

https://blog.datahut.co › web-scraping-at-large-data-extraction-challenges-yo...

Dataset:

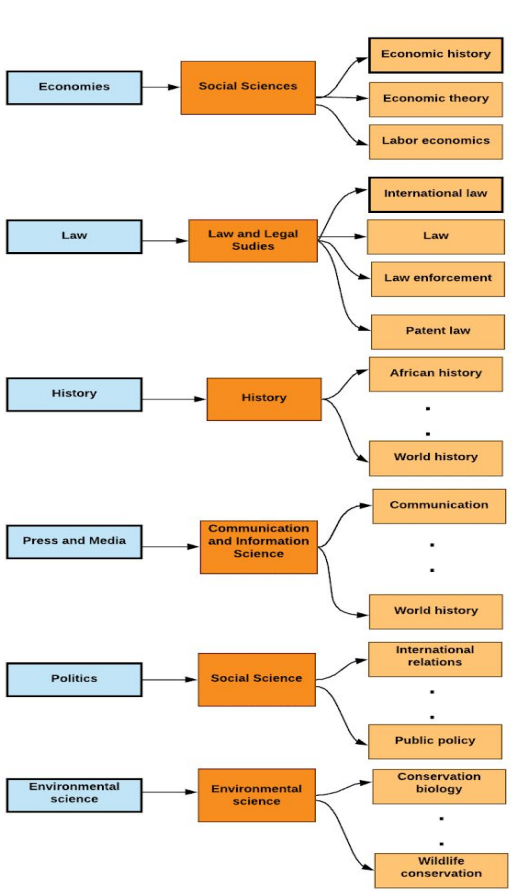
- Saudi Digital Libraries
  - AskZad Library
  - 12 categories
    - Total 518 documents
    - 124,320 words

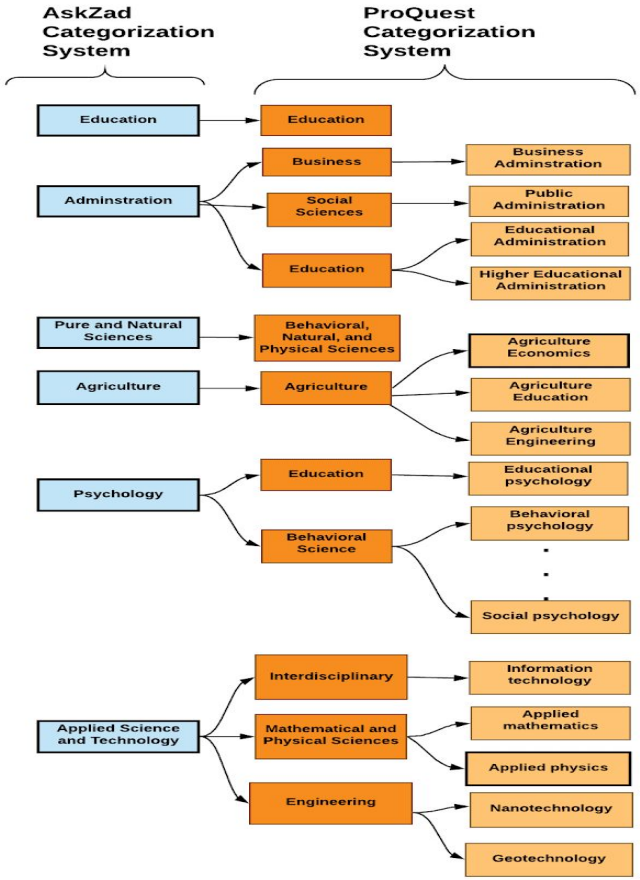


Category	Number of Documents
Education	50
Administration	50
Pure and Natural Sciences	18
Agriculture	29
Psychology	47
Applied Science and Technology	11
Economies	37
Law	50
History	50
Press and Media	78
Politics	80
Environmental Science	18

Categories:

- Mapping to ProQuest categorization system





Preprocessing:

1. Stopwords removal
   a. NLTK
2. Lemmatization
   a. By Farasa API

Lemmatization works better than stemming for the data mining and information retrieval, especially in Arabic as it is highly inflectional language.

Lemmatization

Python 3

JavaScript

Java

```
import http.client

conn = http.client.HTTPSConnection("farasa-api.qcri.org")
payload = "{\"text\": \"هذا مثال بسيط\"}"

headers = { 'content-type': "application/json", 'cache-control':
"no-cache", }

conn.request("POST", "/msa/webapi/lemma", payload, headers)

res = conn.getresponse()

data = res.read()

print(data.decode("utf-8"))
```

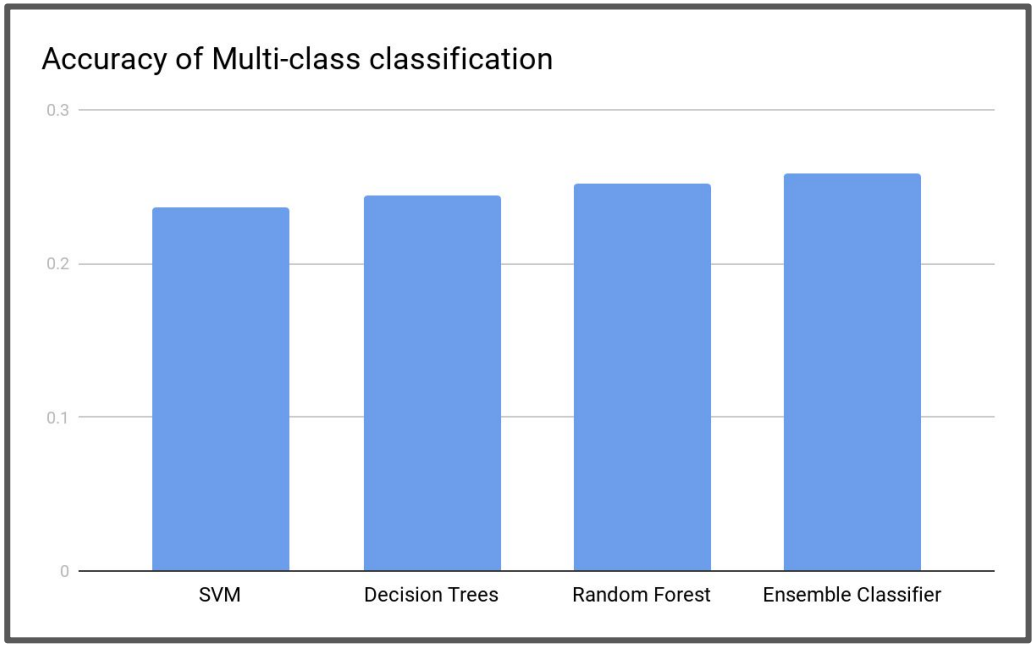
Experiments and Preliminary Results

- Multiclass classification performed poorly:
  - Average Accuracy ~ 24%
- Binary classification performed better:
  - Average Accuracy ~ 68% per Category

Experiments and Preliminary Results (Contd.):

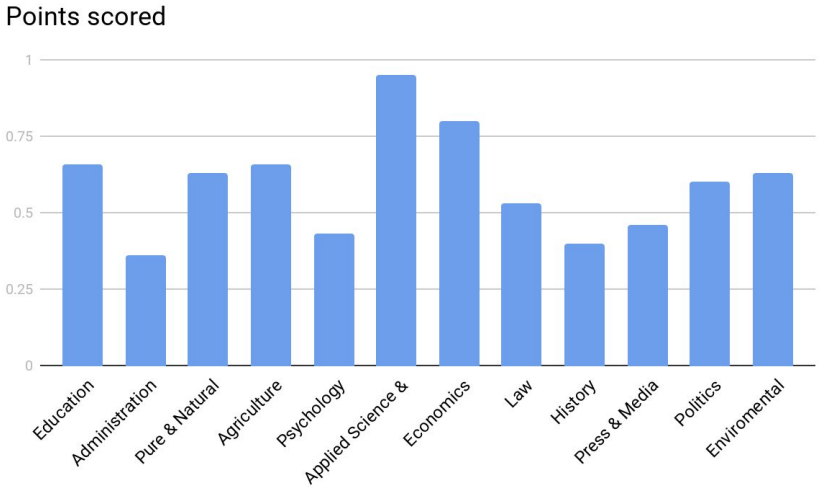
- *Multi-class Classification:*

Classifier	Accuracy
SVM	0.237
Decision Trees	0.244
Random Forest	**0.252**
Ensemble Classifier	0.259



Experiments and Preliminary Results (Contd.):

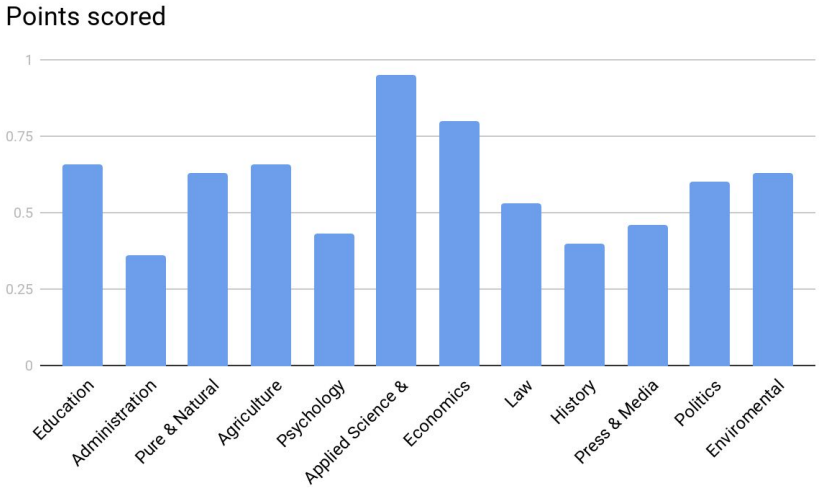
- *Binary Classification:*
  - *Random Forest*



Category	Precision	Recall	F-Measure	Accuracy
Education	0.69	0.66	0.65	0.66
Administration	0.18	0.45	0.26	0.36
Pure and Natural Sciences	0.62	0.63	0.62	0.63
Agriculture	0.69	0.65	0.64	0.66
Psychology	0.21	0.5	0.30	0.43
Applied Science and Technology	0.95	0.95	0.95	0.95
Economies	0.8	0.79	0.79	0.8
Law	0.57	0.57	0.53	0.53
History	0.43	0.43	0.39	0.4
Press and Media	0.42	0.42	0.42	0.46
Politics	0.3	0.37	0.37	0.6
Environmental Science	0.79	0.60	0.54	0.63

Experiments and Preliminary Results (Contd.):

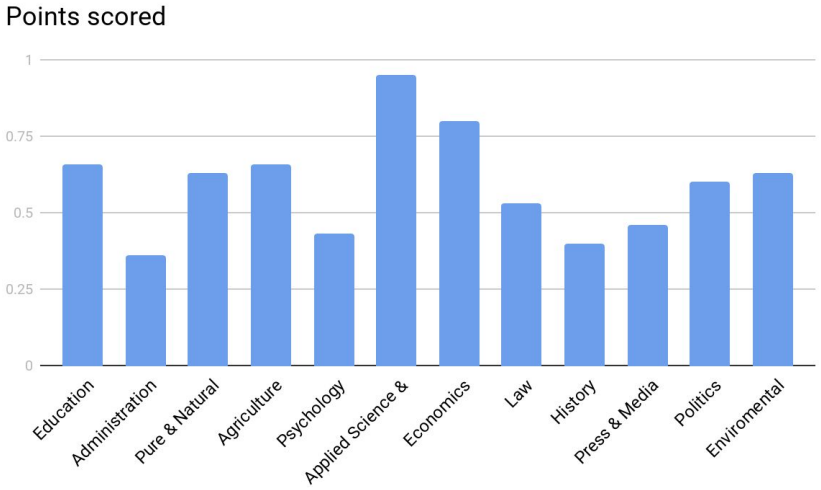
- *Binary Classification:*
  - *Random Forest*



Category	Precision	Recall	F-Measure	Accuracy
Education	0.69	0.66	0.65	0.66
Administration	0.18	0.45	0.26	0.36
Pure and Natural Sciences	0.62	0.63	0.62	0.63
Agriculture	0.69	0.65	0.64	0.66
Psychology	0.21	0.5	0.30	0.43
Applied Science and Technology	0.95	0.95	0.95	0.95
Economies	0.8	0.79	0.79	0.8
Law	0.57	0.57	0.53	0.53
History	0.43	0.43	0.39	0.4
Press and Media	0.42	0.42	0.42	0.46
Politics	0.3	0.37	0.37	0.6
Environmental Science	0.79	0.60	0.54	0.63

Experiments and Preliminary Results (Contd.):

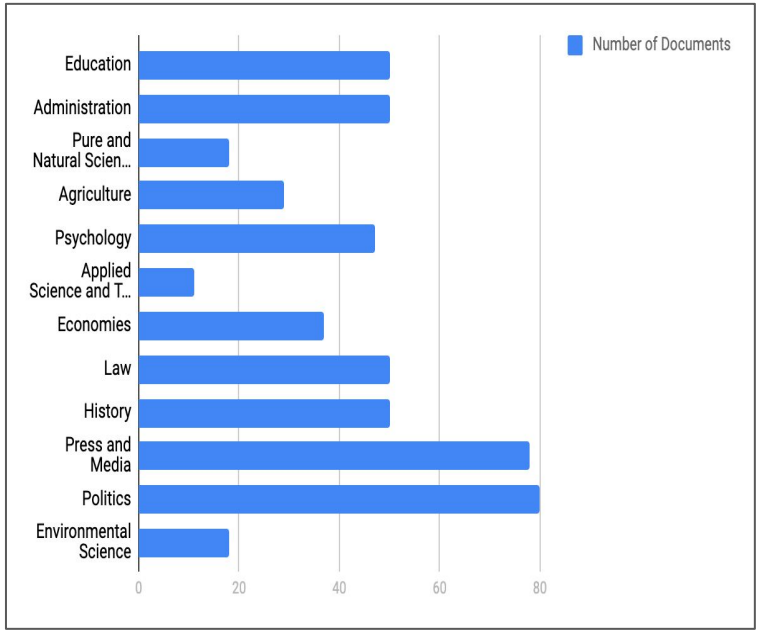
- *Binary Classification:*
  - *Random Forest*



Category	Precision	Recall	F-Measure	Accuracy
Education	0.69	0.66	0.65	0.66
Administration	0.18	0.45	0.26	0.36
Pure and Natural Sciences	0.62	0.63	0.62	0.63
Agriculture	0.69	0.65	0.64	0.66
Psychology	0.21	0.5	0.30	0.43
Applied Science and Technology	0.95	0.95	0.95	0.95
Economies	0.8	0.79	0.79	0.8
Law	0.57	0.57	0.53	0.53
History	0.43	0.43	0.39	0.4
Press and Media	0.42	0.42	0.42	0.46
Politics	0.3	0.37	0.37	0.6
Environmental Science	0.79	0.60	0.54	0.63

Insights and Future work

- Investigate why there exists a big difference between accuracies for different categories in the Binary Classification.



Category	Precision	Recall	F-Measure	Accuracy
Education	0.69	0.66	0.65	0.66
Administration	0.18	0.45	0.26	0.36
Pure and Natural Sciences	0.62	0.63	0.62	0.63
Agriculture	0.69	0.65	0.64	0.66
Psychology	0.21	0.5	0.30	0.43
Applied Science and Technology	0.95	0.95	0.95	0.95
Economies	0.8	0.79	0.79	0.8
Law	0.57	0.57	0.53	0.53
History	0.43	0.43	0.39	0.4
Press and Media	0.42	0.42	0.42	0.46
Politics	0.3	0.37	0.37	0.6
Environmental Science	0.79	0.60	0.54	0.63

Insights and Future work

- Investigate why there exists a big difference between accuracies for different categories in the Binary Classification.
- Investigate the low performance of the Multi-class Classification:
  - Parameters tuning

Insights and Future work

- Investigate why there exists a big difference between accuracies for different categories in the Binary Classification.
- Investigate the low performance of the Multi-class Classification:
  - Parameters tuning
- Increase the size of the corpus:
  - Sketch Engine

Insights and Future work

- Investigate why there exists a big difference between accuracies for different categories in the Binary Classification.
- Investigate the low performance of the Multi-class Classification:
  - Parameters tuning
- Increase the size of the corpus .
  - Sketch Engine
- Run each classifier against both Arabic and English abstracts separately.
- Use word embeddings.


Questions