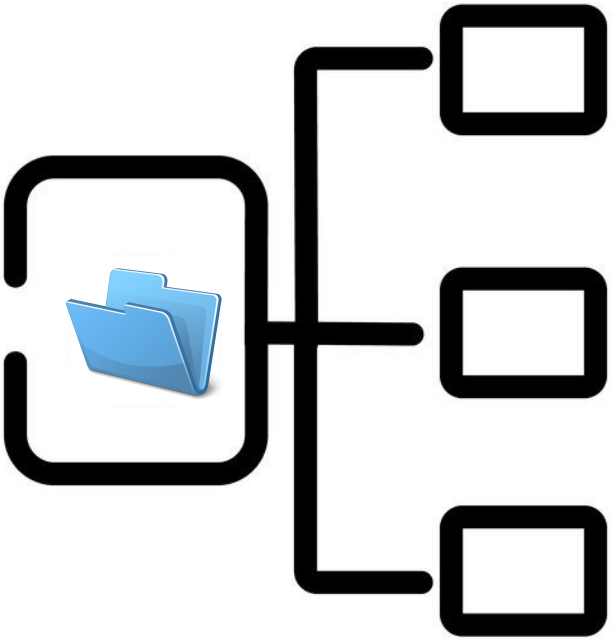


Supervised Learning for Arabic Text Classification to Organize the Content of UN Platforms

This paper has been prepared by Asmaa Ali, Technology Trainee at the ESCWA Technology Center.

February 2021

Contents

Supervised Learning

The type of learning used in the majority of machine learning applications is the supervised learning where there is an input data (x) and output labels (y) and machine learning algorithms are used to learn how the mapping from the input to the output occurs.

$$\mathbf{Y = f(X)}$$

The learned mapping is then used for predicting the output of any new input. It's called supervised learning because there are output labels available that work as a teacher supervising the learning process. Learning is an iterative process and it stops when the algorithm achieves an acceptable performance.

Problem Definition

The main goal of this project is to provide a model that will be able to classify any arabic content to be used for two United nations Platforms, MSME (micro, small, medium enterprises and entrepreneurship) and DIAR (Driving the innovation in the Arab Region Science, Technology & Innovation). There are three categories in the dataset used for training, entrepreneurship, Science & Technology and Other.

Data Collection

For the previously mentioned three categories, around 2800 records for each category were used for models training and testing. Here are the sources of the data collected for each category:

- The Entrepreneurship Data was collected from eight websites: ryadibusiness, waya, youm7, jawlah, asharqbusiness, raedaamal, egyentrepreneur, and preneur-masr.
- The Science & Technology Data was collected from three websites: RT-Online, asharq, and sputniknews.
- The Data of the Other category was collected from three websites: UN News, birzeit university, and almashareq.

Text Pre-processing

The goal of this phase is to prepare the data for training through unifying the data form to be passed to the model. Here are the four steps were used for data preprocessing:

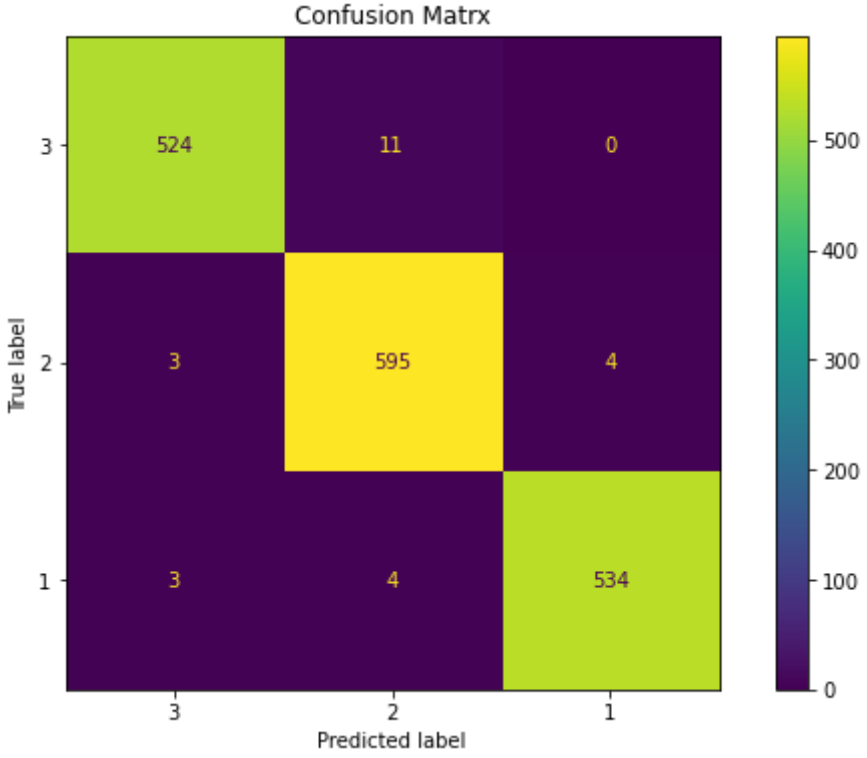
- Removing punctuations
- Removing arabic diacritics.
- Removing longation.
- Removing stop words.

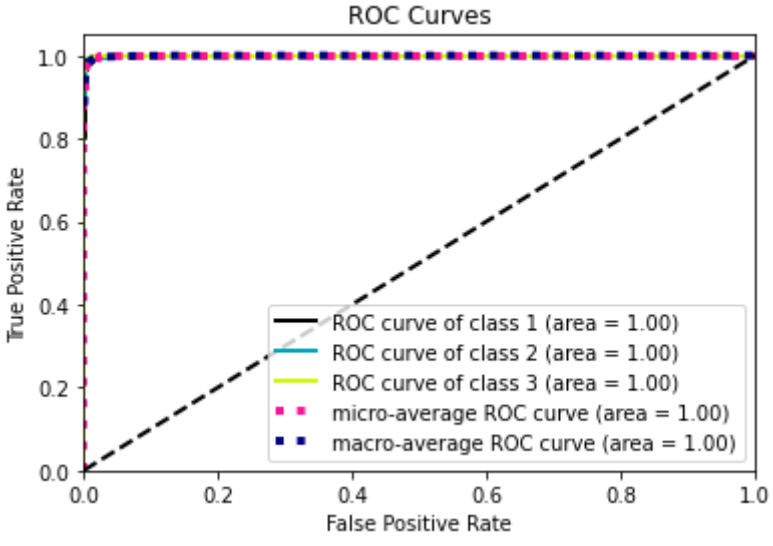
Results

Logistic Regression

Accuracy score is 0.99

	precision	recall	f1-score	support
1	0.99	0.98	0.98	535
2	0.98	0.99	0.98	602
3	0.99	0.99	0.99	541
accuracy			0.99	1678
macro avg	0.99	0.98	0.99	1678
weighted avg	0.99	0.99	0.99	1678

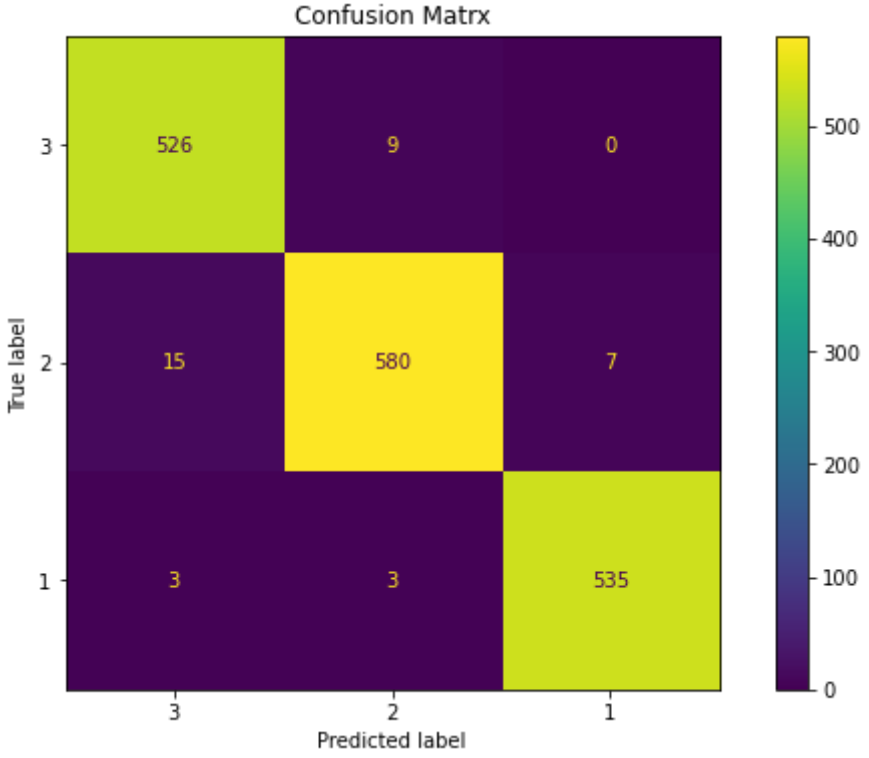


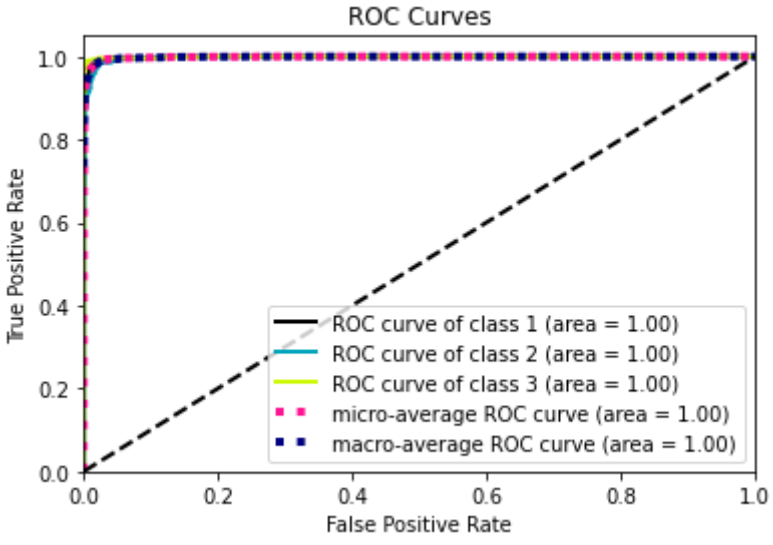


Random Forest

Accuracy score is 0.98

	precision	recall	f1-score	support
1	0.97	0.98	0.97	535
2	0.98	0.96	0.97	602
3	0.99	0.99	0.99	541
accuracy			0.98	1678
macro avg	0.98	0.98	0.98	1678
weighted avg	0.98	0.98	0.98	1678

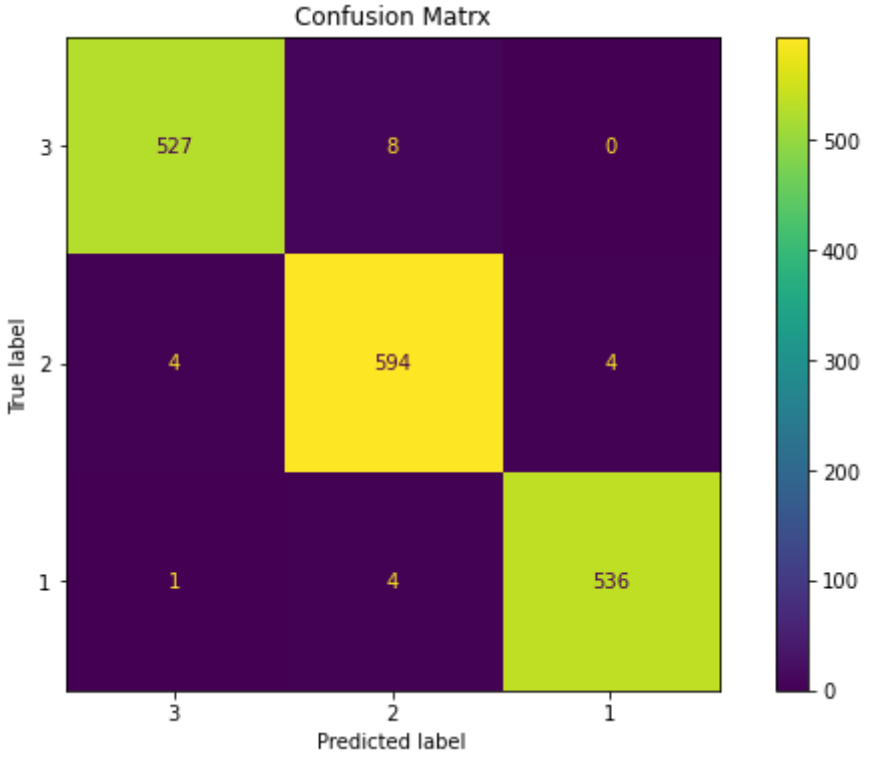




Support Vector Machines

Accuracy score is 0.99

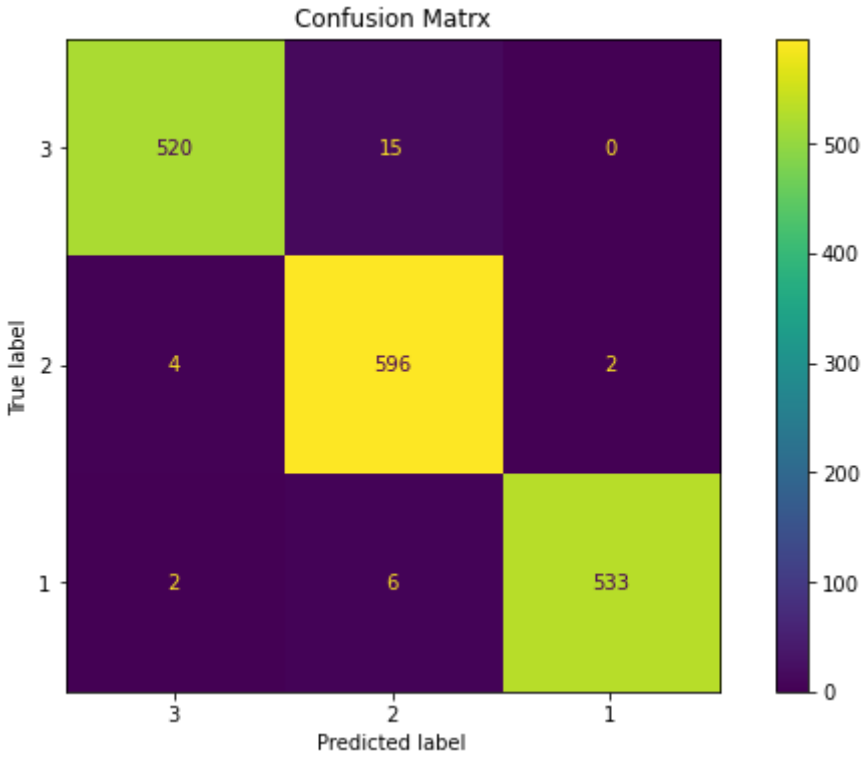
	precision	recall	f1-score	support
1	0.99	0.99	0.99	535
2	0.98	0.99	0.98	602
3	0.99	0.99	0.99	541
accuracy			0.99	1678
macro avg	0.99	0.99	0.99	1678
weighted avg	0.99	0.99	0.99	1678



Stochastic Gradient Descent

accuracy 0.9827175208581644

	precision	recall	f1-score	support
1	0.99	0.97	0.98	535
2	0.97	0.99	0.98	602
3	1.00	0.99	0.99	541
accuracy			0.98	1678
macro avg	0.98	0.98	0.98	1678
weighted avg	0.98	0.98	0.98	1678



Conclusion

Results show that SVM achieved the best performance on arabic content classification, outperforming the other methods.