Achtatou Sean
Siwen Guo

University of Luxembourg

Bachelor in Computer Science

The Machine Learning
for sentiment analysis
(Short Version)

Semester 3 - 2018-2019

1.Introduction

This research work, for a semester project in the Bachelor in Computer Science in the University of Luxembourg, is about a field of the computer science which is getting important for the future of the humanity, since the last several years. This important field, which is evolving and growing days after days, is nothing else than the Artificial Intelligence. Indeed, for this project, we focused on one of its sub-domain, which is the Machine Learning. The Machine Learning is an Artificial Intelligence which can be used for the classification of big data of information by the use of a computer.

2.Project Description

2.1. Domain

The domain of this project concerns one of the many interesting domains of the computer science: The Artificial Intelligence. With the technology of implementation of intelligence in computers evolving days after days, the Artificial Intelligence has become one of the main sector to work and study with in the computer science.

The Artificial Intelligence is mainly a technology allowing the humanity to work more efficiently and faster on basic tasks, thanks to its capabilities and power of calculation the humans don't own. The Artificial Intelligence may replace step by step multiples human's activities and provides a faster evolution in the humanity societies.

2.2. Objectives

The main objective of this project is to get to know with the operation of an Artificial Intelligence. Indeed, most of the people believe an Artificial Intelligence as a machine which is really intelligent. However, it is not the case at all. Indeed, with this project we will the possibility to show that an Artificial Intelligence does mostly operate on calculation of probabilities, analysis and statistics.

The second objective, which is the continuation of the main objective, is to provide to the mankind a trustworthy regarding the Intelligent systems in their future daily life. Since, there may have an eventuality that the Intelligent systems will evolve around the population.

2.3. Constraints

The first constraint of this project is the programming background. Since, we will need to implement a Machine Learning on a programmation language, the notion of programmation may not have to be unknown. For this project we will use the Python 3 language beginning to be widely used for the Artificial Intelligence.

The last constraint is the basics mathematical background. Indeed, the notion of operation of an Artificial Intelligence is difficult to understand at first sight. Some knowledge in mathematics, such as the probabilities and statistics have to be known to be able to fully understand the conception of the Machine Learning.

3.Background

3.1. Scientific

3.1.1. The Artificial Intelligence

The Artificial Intelligence is a program aiming to predict multiples events based on its actual knowledge and analyzing, to produce a satisfactory result. The Artificial Intelligence can be used for multiples tasks in different domains such as:

- Analyzing words/images
- Manipulation of objects
- Driving cars
- Conversation with humans
- …

3.1.2. Types of learning

The Artificial Intelligence can be used with three different types of learning:

- Supervised
- Unsupervised
- Semi-supervised

The supervised learning, which is being used for this project, the Artificial Intelligence will use data already classified.

The Unsupervised learning, the Artificial Intelligence will use data not classified.

The semi-supervised learning, is a mix of supervised and unsupervised learning.

3.1.3. Machine Learning versus Deep Learning

The Machine learning is based on the building of models with big set of data, which can be used later to predict possible new data.

The Deep learning is more based on neural networks, on the learning of data representation, meaning it has to discover itself how to use the given data.

3.2. Technical

3.2.1. The Support Vector Machine

The Support Vector Machine is a Supervised learning technique, creating models to be used for the classification of data. It used a part of the data as training set to create the model, and the other part as testing data to test the accuracy and power of prediction of the models.

3.2.2. Introduction to pre-processing

The pre-processing is a step in which the data is modified in order to be used by the Support Vector Machine to create the models. Involving in multiples steps:

Split sentence in tokens
Get rid of special characters
Split the sentences in words
Discard the stop words
Lemmatize the tokens

Remaining tokens will be features.

3.2.3. The Glove library

The Glove library is an external library allowing us to create a set of vectors for the data based on their own features. It results in obtaining a set of fixed length of vectors being less long, more efficient and consequently providing a better accuracy, to input in the models.

4.Results

For this project, we used a supervised Machine Learning with the Support Vector Machine on big set of messages to predict a possible good or bad sentiment. We worked on two corpuses, the first focusing on global messages, the second focusing on terrorism behaviour meaning we had to sort the messages first. For each corpus we provided two different models, linear and kernel, allowing us to work in three dimensional model. Moreover, for the second corpus we used the Glove library for the two models. For each of the models, we calculated the accuracy and the power of prediction to observe their capability to provide a satisfactory result.

For the first model being linear, not using the Glove, with the first corpus. We obtained an accuracy of prediction of eighty-two. Moreover, we obtained as power of prediction the following table.**[Picture 1]**This table is the representation of the percentage of correct prediction of a good or a bad behaviour.

For the second model using the kernel, not using the Glove, with the first corpus. We obtained an accuracy of prediction of eighty-six. As observed, we obtained an accuracy higher than for the last model thanks to the use of the kernel. Moreover, we obtained as power of prediction the following table. **[Picture 2]**

For the third model being linear, using the Glove, with the second corpus. We obtained an accuracy of prediction of sixty-one. This result can be explaining by the fact we had to sort the messages having a possible relation to terrorism behaviour, since it is done manually it can bring noise to the result. However, when we observe the table of prediction. **[Picture 3]**

For the last model using the kernel, using the Glove, with the second corpus. We obtained an accuracy of prediction of sixty-four. Once more, the use of the kernel still increases the accuracy of prediction for the model. Moreover, when we observe the table of prediction.
[Picture 4]

5.Conclusion

In conclusion, we could observe with the results obtained from the individual models, created with the Machine Learning using support vector machine, that our models have the possibility to be used in a real situation to classify messages according to the sentiment or terrorism behaviour. Moreover, we could note that by using the kernel, the models had a better power of prediction than without it.

In last, we could see that an Artificial Intelligence operation is simply based a set of probabilities and analysis, but can not be considered as a hundred percent trustworthy since it can still raise an error.

Appendix

Prediction/Sentiment	Positive	Negative
True	93 %	89 %
False	7 %	11 %

Picture 1. Results for the table prediction of the model linear with the corpus "corpus-medium.txt" without using Glove.

Prediction/Sentiment	Positive	Negative
True	96 %	98 %
False	4 %	2 %

Picture 2. Results for the table prediction of the model using the kernel trick with the corpus "corpus-medium.txt" without using Glove.

Prediction/Sentiment	Positive	Negative
True	89 %	72 %
False	11 %	28 %

Picture 3. Results for the table prediction of the model linear with the corpus "corpus-1m600.txt" using Glove.

Prediction/Sentiment	Positive	Negative
True	93 %	96 %
False	7 %	4 %

Picture 4. Results for the table prediction of the model using the kernel trick with the corpus "corpus-1m600.txt" using Glove.