# **Sentiment Analysis of Movie Reviews**

Natural Language Processing Project

# Team 29

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Id | Department | Level |
| احمد هاني احمد عبدالمجيد | 20201700090 | SC | 4 |
| أحمد وائل صلاح الدين حسن | 20201700091 | SC | 4 |
| جورج سعيد سليم ساويرس | 20201700210 | SC | 4 |
| جنه ايمن طلعت مصطفي | 20201700204 | SC | 4 |
| اسراء سعد عبدالحميد سعيد | 20201701055 | SC | 4 |
| عبدالرحمن محمد عزازي محمد | 20201700483 | SC | 4 |

1. Data Augmentation:

Double the size of the used dataset by back translation. It is done by translating each review in the dataset from English to French, then translating it back to English using google translator Api.

The default dataset was 1000 positive review and 1000 negative review, but after data augmentation it increased to 2000 positive review and 2000 negative review.

1. Preprocessing:

Several techniques of preprocessing where used:

* Stop words removal.
* Word tokenization.
* Word Lemmetization.
* Remove punctuation.

1. Feature Extraction:

Term frequency – Inverse document frequency (TF-IDT) algorithm was used to extract features from the reviews.

1. Data Splitting:

The used dataset was divided into 80% for training and 20% for testing.

1. Grid Search:

Grid Search was applied on three machine learning classification models:

* **Logistic Regression :**

Best Accuracy is 96% using hyperparameters (c : 100, max\_iter : 100, solver : ‘newton-cg’)

* **Support Vector Machine :**

Best Accuracy is 94.5% using hyperparameters (c : 10, gamma : 0.1, kernel : ‘rbf’)

* **Random Forest :**

Best Accuracy is 91% using hyperparameters (max\_depth : 30, min\_samples\_leaf : 2, min\_samples\_split : 2, n\_estimators : 300)

1. Model Training And Evaluating:

We trained four machine learning classification models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-score |
| Logistic Regression | 96% | 96.5% | 96% | 96% |
| Support Vector Machine | 94.5% | 94% | 95.5% | 95% |
| Random Forest | 91% | 93% | 92% | 91.5% |
| Naïve Bayes | 89% | 92% | 86% | 89% |

1. Results Visualization:

