Big Data in Media Technology

Lab 1 : Sentiment Classification with Naïve Bayesian Classifier

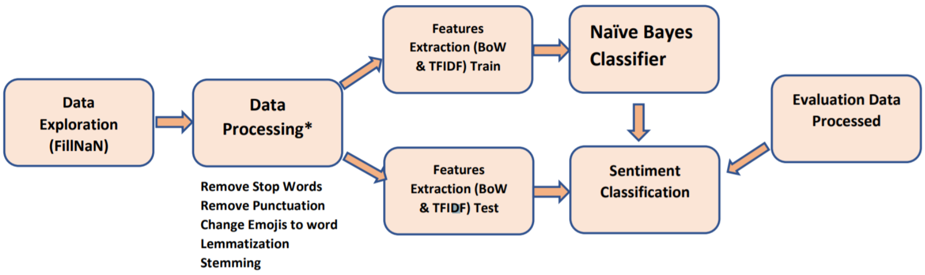
Team 2

Hamza AJJA

Yassine ELKHEIR

3^{rd} Year - 15 septembre 2021

I Analysis Process Diagram :



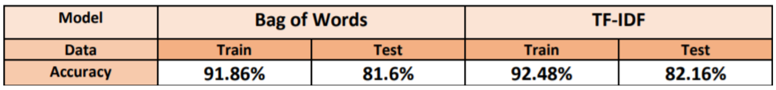
II Features Extraction :

1-Bag Of Words : A bag-of-words model is a way of extracting features from text for use in modeling, such as with machine learning algorithms. The approach is very simple and flexible and can be used in a myriad of ways of extracting features from documents. A bag-of-words is a presentation of text that describes the occurrence of words within a document. It involves two things :

1. A vocabulary of known words.
2. A measure of the presence of known words.

2-Term Frequency – Inverse Document Frequency (TF-IDF) : In this model, the words are assigned a weight based on the frequency of appearance. The model has 2 parameters. The term frequency component adjusts the weight proportionally with the number of times the word appears in the document with respect to the total number of words in that document. Inverse document frequency component identifies unique words in the set of documents and increases weight accordingly.

III Evaluation : (Metric Used = Accuracy) :

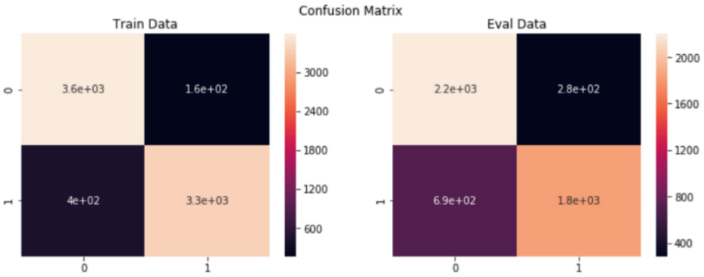


Model	Bag of Words		TF-IDF	
Data	Train	Test	Train	Test
Accuracy	91.86%	81.6%	92.48%	82.16%

After analyzing the results found with the two models Bag of Words and TF-IDF, we take the model that gives us the highest accuracy (we can see that both models suffer from the overfitting problem, with another more sophisticated model like Support Vector Machine, we can achieve better results)

IV Results :

We have reached an accuracy of 80.58% on Evaluation Data. We aim to develop our model by try some sophisticated classification models such as SVM, Neural Networks, to get a higher accuracy, and well tune parameters to avoid the overfitting problem .


Confusion Matrix
Train Data
3.6e+03
1.6e+02
4e+02
3.3e+03
Eval Data
2.2e+03
2.8e+02
6.9e+02
1.8e+03